DIEGO BERTOLINI GONÇALVES

# IDENTIFICAÇÃO E VERIFICAÇÃO DE ESCRITORES USANDO CARACTERÍSTICAS TEXTURAIS E DISSIMILARIDADE

Tese apresentada ao Programa de Pós-Graduação em Informática do Setor de Ciências Exatas da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutor em Ciência da Computação.
Orientador: Luiz Eduardo S. Oliveira, Dr.
Co-orientador: Robert Sabourin, Dr.

CURITIBA - PR

2014

G635i Gonçalves, Diego Bertolini.

Identificação e verificação de escritores usando características texturais e dissimilaridade/ Diego Bertolini Gonçalves. – Curitiba, 2014.
117f. : il. [algumas color.]; 30 cm.

Tese (Doutorado) - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-graduação em Informática, 2014.

Orientador: Luiz Eduardo S. Oliveira. Co-Orientador: Robert Sabourin.
Bibliografia: p. 109-117.

1. Escritores. 2. Escrita. I. Oliveira, Luiz Eduardo S. II. Sabourin, Robert. III. Universidade Federal do Paraná. IV. Título.

CDD: 808.8

Ministério da Educação
Universidade Federal do Paraná
Programa de Pós-Graduação em Informática

## PARECER

Nós, abaixo assinados, membros da Banca Examinadora da defesa do aluno de Doutorado em Ciência da Computação, Diego Bertolini Gonçalves, avaliamos a tese de doutorado intitulada *"Identificação e verificação de escritores usando características texturais e dissimilaridade"*, cuja defesa pública foi realizada no dia 07 de julho de 2014, às 14:00 horas, no Departamento de Informática do Setor de Ciências Exatas da Universidade Federal do Paraná. Após avaliação, decidimos pela:

(X) **aprovação** do candidato. ( ) **reprovação** do candidato.

Curitiba, 07 de julho de 2014.

Prof. Dr. Luiz Eduardo Soares de Oliveira
**DINF/UFPR** – Orientador

Prof. Dr. Helio Pedrini
**UNICAMP** – Membro Externo

Prof. Dr. Edson Justino
**PUC-PR** – Membro Externo

Profa. Dra. Giselle Ferrari
**Elétrica/UFPR** – Membro Externo

Prof. Dr. Lucas Ferrari de Oliveira
**DINF/UFPR** – Membro Interno

# AGRADECIMENTOS

Gostaria de agradecer a todos que direta ou indiretamente colaboraram para elaboração deste trabalho. Amigos, professores, colaboradores, revisores, entre tantas pessoas. Em especial gostaria de agradecer:

Aos meus pais, Sérgio e Marlene a quem devo tudo que sei e tudo que sou. Pai, mãe, serei eternamente grato por tudo que fizeram por mim.

A minha esposa Caroline, que desde o início me incentivou, compreendeu minhas ausências, me apoiou nas horas difíceis e sempre esteve ao meu lado.

Ao meu orientador Luiz E. S. Oliveira, o qual com paciência e sabedoria soube muito bem desempenhar a função de orientador, professor e amigo. Que compreendeu minhas dificuldades e minhas ausências. Luiz, sou muito grato a você por me incentivar, por tudo que me ensinou e pelos trabalhos desenvolvidos em parceria. Torço para que possamos levar esta parceria adiante.

Ao Dr. Robert Sabourin pelo suporte e contribuições em todos estes anos.

Aos professores, Dr. Edson José Rodrigues Justino, Dr. Helio Pedrini, Dr. Giselle Ferrari, e Dr. Lucas Ferrari. Seus comentários ajudaram a melhorar a versão final da tese.

Aos amigos, em especial a Eduardo Sant´Ana, Jefferson Martins, Yandre Maldonado e Pedro Luiz de Paula.

A todos os professores da UTFPR - Campo Mourão, em especial a Juliano Foleiss pela ajuda, comentários e interesse no meu trabalho.

A toda minha família pela força, em especial ao meu irmão Sérgio Júnior, minha avó Ezia Marroni (*in memorian*) e meu avô Anésio Bertolini.

A Deus, por ter me amparado nos momentos difíceis e por ter me guiado em mais esta caminhada.

# RESUMO

A verificação e identificação de escritores são atividades relacionadas à ciências forense, na qual possuem à função de auxiliar na identificação ou constatação de fraudes de documentos manuscritos. A tarefa de verificar ou identificar escritores através de sua escrita manuscrita disposta em papel torna-se árdua devido às semelhanças existentes entre a escrita de diferentes escritores e também devido a variabilidade da escrita de uma mesma pessoa. Inserido neste contexto, este trabalho discute o uso de descritores de textura para o processo de verificação e identificação de escritores. Três diferentes descritores de textura foram avaliados para elaboração desta tese, GLCM (*Gray Level Co-occurrence Matrix*), LBP (*Local Binary Pattern*) e LPQ *Local Phase Quantization*. Além disso, empregamos um esquema de classificação baseado na representação da dissimilaridade, o qual tem contribuído para o sucesso em problemas de verificação de escritores. Inicialmente tratamos de algumas questões, como o desempenho dos descritores e parâmetros do sistema escritor-independente. Observamos outras questões importantes relacionadas com a representação dissimilaridade, tais como o impacto do número de referências utilizadas para verificação e identificação de escritores, e o número de escritores empregados no conjunto de treinamento. A partir destes primeiros experimentos, foi possível verificar que o número de escritores no conjunto de treinamento impactava menos que se supunha no desempenho do sistema. Para verificar todos estes objetivos, realizamos experimentos com duas diferentes bases de dados: BFL (*Brazilian Forensic Letter Database*) e IAM (*Institut für Informatik und angewandte Mathematik*), as quais são manuscritas em diferentes línguas e contendo números de escritores díspares. Em sequência, comparamos a abordagem baseada na dissimilaridade com outras estratégias escritor-dependente. Em uma segunda etapa de experimentos avaliamos o impacto de diferentes estilos de escrita, assim como: texto-dependente, texto-independente, caixa alta e falsificação (escrita dissimulada). Para isso, utilizamos a base *Firemaker* a qual é a única base pública a possuir estes quatro diferentes estilos. Por fim avaliamos a abordagem de seleção de escritores a qual tem por finalidade selecionar escritores para geração de modelos robustos. Através de uma série de experimentos, percebemos que ambos os descritores de textura LBP e LPQ são capazes de superar os resultados anteriores descritos na literatura para o problema de verificação por cerca de 5 pontos percentuais. Para o problema de identificação de escritores, o uso do descritor LPQ foi capaz de alcançar melhores taxas de acertos globais, 96,7 % e 99,2 % para as bases BFL e IAM, respectivamente. Com relação aos diferentes estilos de escrita, notamos que a abordagem apresenta-se robusta para diferentes estilos incluindo a falsificação, apresentando desempenho superior aos descritos em literatura. Por fim, utilizando a abordagem de seleção de escritores, foi possível alcançar desempenho igual ou superior utilizando cerca de 50% dos escritores disponíveis no conjunto de treinamento.

**Palavras-Chave:** Identificação de escritores, reconhecimento de padrões, dissimilaridade, textura, seleção de escritores.

# ABSTRACT

Writer verification and identification are related to forensic science activities, which are used in the tasks of identifying or finding fraud in handwritten documents. The task of writer verification or identification through handwriting in paper becomes difficult due to similarities between the writing of different writers and also because of the variability of the handwriting of a given person. In this context, this work discusses the use of texture descriptors for the verification process and identification of writers. Three different texture descriptors were considered in this study, GLCM (Gray Level Co-occurrence Matrix), LBP (Local Binary Pattern) and LPQ (Local Phase Quantization). A classification scheme based on dissimilarity representation, which has contributed to the success in writer verification problems, was adopted in this work. Initially we addressed some issues, such as performance of descriptors and parameters of the writer-independent system. We have also observed other important issues related to the dissimilarity representation, such as the impact of the number of references used for verification and identification of writers, and the number of writers employed in the training set. From these initial experiments, we found that the number of writers in the training set has low impact in system performance. To accomplish all of these goals, we conducted experiments on two different databases: BFL (Brazilian Forensic Letter Database) and IAM (Institut fur Informatik und angewandte Mathematik), which are acquired in different languages containing different numbers of writers. Next, we compared the scheme based on dissimilarity representation with other writer-dependent approach strategies. In a second round of experiments we evaluated the impact of different writing styles, as well as text-dependent, text-independent, upper case and forgery (disguised writings). To that end, we used the Firemaker database which is the only public database that has these four different styles. Finally, we proposed an approach for selecting writers to build a better dissimilarity model. Through a series of experiments, we noticed that both texture descriptors LBP and LPQ are able to outperform previous results reported in the literature for the problem of verification by about 5%. Regarding the problem of writer identification, the LPQ descriptor was able to achieve better identification rates for global hits, 96.7% and 99.2% for IAM and BFL databases, respectively. With respect to the different styles of writing, we have shown that the approach is robust for different styles including forgery, presenting higher performance than those described in literature. Finally, using the proposed writer selection method, it was possible to achieve equal or better performance using about 50% of writers available in the training set.



**Keywords:** Writer identification, pattern recognition, dissimilarity, texture, writers selection.

# LISTA DE FIGURAS

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS

| | |
|---|---|
| AR | Auto Regressivo |
| ALVOT | Algoritmos de Voto |
| AUC | *Area Under a ROC Curve* |
| BFL | *Brazilian Forensic Letter Database* |
| CEDAR | *Center of Excellence for Document Analysis and Recognition* |
| CMC | *Cumulative Match Curve* |
| CENPARMI | *Centre for Pattern Recognition and Machine Intelligence* |
| DFT | *Discrete Fourier Transform* |
| DHT | *Discrete Hermite Transform* |
| DWT | *Discrete Wavelet Transform* |
| EER | *Equal Error Rate* |
| FDP | *Probability Density Function* |
| FN | *False Negative* |
| FP | *False Positive* |
| GLCM | *Gray Level Co-occurrence Matrix* |
| GLRL | *Gray Level Run Length* |
| GMM | *Generalized Markov Model* |
| HMM | *Hidden Markov Model* |
| IAM | *Institut für Informatik und angewandte Mathematik* |
| ICA | *Independent Component Analysis* |
| ICDAR | *International Conference on Document Analysis and Recognition* |
| ICFHR | *International Conference on Frontiers in Handwriting Recognition* |
| ISI | *Handwritten Character Database of India Scrits* |
| $k$-NN | *k-Nearest Neighbors* |
| LBP | *Local Binary Pattern* |
| $\text{LBP}^{ri}$ | LBP *with rotation invariant* |
| $\text{LBP}^{riu2}$ | LBP *with rotation invariant uniform 2 pattern code* |
| $\text{LBP}^{u2}$ | LBP *with uniform 2 pattern code* |
| LBSR | *Line-Based Spectrum Resolution* |
| LPQ | *Local Phase Quantization* |
| MC | *Confusion Matrix* |
| MCD | *Multichannel Decomposition* |
| MCS | *Multiple Classifier Systems* |
| MLP | *Multilayer Perceptron* |
| MSER | *Maximally Stable Extremal Regions* |
| NDDF | *Normal Density Discriminant Function* |
| NIR | *Near-Infrared* |
| NIST | *National Institute of Standards and Technology* |
| oBIF | *Basic Image Feature Columns* |
| PCA | *Principal Component Analysis* |

Continua na próxima página

continuação da página anterior

| | |
|---|---|
| QUWI | *Qatar University Writer Identification dataset* |
| RIMES | *Reconnaissance et Indexation de données Manuscrites et de fac similÉS / Recognition and Indexing of handwritten documents and faxes* |
| RNA | *Artificial Neural Network* |
| RBF | *Radial Basis Function* |
| ROC | *Receiver Operator Characteristics* |
| SIFT | *Scale Invariant Feature Transform* |
| SOM | *Self-Organizing Maps* |
| STFT | *Short-Term Fourier Transform* |
| SVM | *Support Vector Machine* |
| SURF | *Speed-Up Robust Feature* |
| TD | *Texto-dependente* |
| TI | *Texto-independente* |
| TEXTEL | *TEXture ELement* |
| TN | *True Negative* |
| TP | *True Positive* |
| VSM | *Vector Space Model* |
| WED | *Weighted Euclidean Distance* |

# SUMÁRIO

# CAPÍTULO 1

# INTRODUÇÃO

O número de aplicações, com intuito de reconhecer padrões, tem crescido rapidamente nas últimas décadas. Este fato se deve aos avanços computacionais e também da necessidade de sistemas com tal capacidade. Em grande parte destes sistemas se almeja alcançar desempenho similar ou até mesmo superior aos apresentados por humanos, já que demonstram grande habilidade no reconhecimento de tais padrões. Assim, sistemas automáticos de reconhecimento de padrões podem contribuir nos mais diversos campos como: industrial, médico, comercial, biométrico, entre outros. Sistemas biométricos vêm sendo pesquisados há décadas e abrangem áreas como o reconhecimento de digitais, reconhecimento de face, reconhecimento da íris, verificação de assinaturas, entre outras. Esta tese tem seu foco na verificação e identificação de escritores através de textos manuscritos.

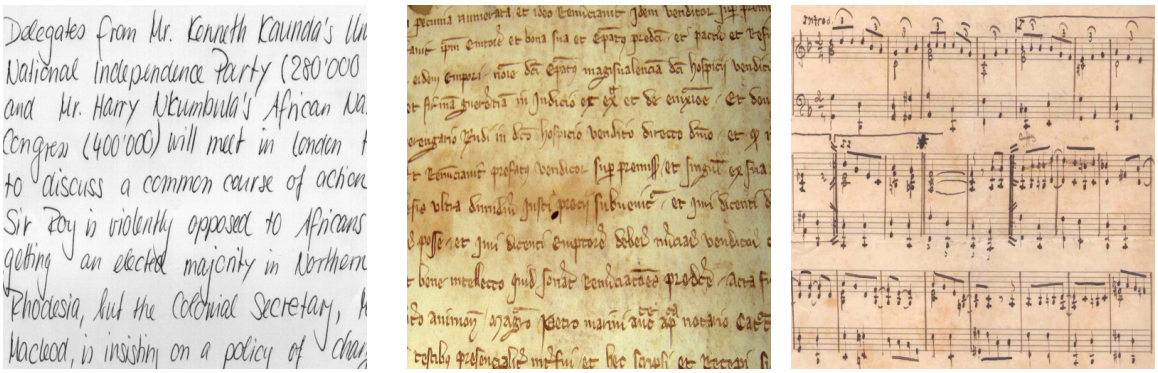
## 1.1 Delimitação do Tema

A primeira verdade com relação aos manuscritos é a impossibilidade de pessoas diferentes possuírem manuscritos idênticos. A segunda verdade é que duas amostras de um mesmo escritor nunca serão exatamente iguais, pois existem variações naturais e pequenos desvios de comportamento na escrita que faz o manuscrito ser uma espécie de assinatura do indivíduo [54].

A escrita, tal como a assinatura, é carregada de características físicas, mentais e emocionais. Detalhes do traçado da escrita revelam um traço da personalidade específica do escritor [91]. Ao observarmos um texto escrito a próprio punho, percebemos que existem muitas características implícitas sobre o escritor. Verificamos, ainda, que existem características únicas que podem ser utilizadas de maneira a distinguir um indivíduo de outro.

Existem diversas linhas de pesquisas correlatas empregando documentos manuscritos, tais como: reconhecimento de texto, reconhecimento de gráficos, análise de manuscrito, reconhecimento de caracteres, verificação de assinaturas, reconhecimento de autoria, além da identificação e verificação de escritor. Sistemas de reconhecimento de autoria através de documentos antigos ou partituras musicais são investigados há anos [67, 14]. A Figura 1.1 apresenta amostras de documentos utilizados no reconhecimento de autoria em três diferentes campos. Nestes casos, temos típicos problemas de classificação, diferenciados apenas pela forma como são representados os documentos do autor.

Anil et al. [49] descrevem três diferentes finalidades para sistemas biométricos com base na escrita manuscrita: identificação, verificação e monitoramento:



(a) (b) (c)

Figura 1.1: Amostras de diferentes campos de pesquisa com identificação de autoria:(a) Manuscritos, (b) Documentos antigos e (c) Partituras.

- **Identificação:** (O escritor está na base de dados?). Refere-se ao processo de recuperação de amostras a partir de uma base de dados de manuscritos. Através da amostra de um manuscrito questionado, desejamos fazer uma consulta deste escritor, fornecendo um subconjunto de documentos relevantes, disponíveis na base de dados, em que análises complementares podem ser realizadas por peritos. Ou seja, determinar o escritor de um manuscrito a partir de um conjunto de escritores, envolvendo uma busca de um para muitos (*1:N*) na base de dados.

- **Verificação:** (O escritor é realmente quem ele afirma ser?). O processo de verificação basicamente classifica uma amostra como sendo genuína ou não genuína. Dado um manuscrito questionado, comparando-o com outros exemplos, desejamos saber se este é um exemplo genuíno (escrito a próprio punho) ou uma falsificação. Ou seja, saber se o manuscrito foi escrito por certa pessoa ou não, neste caso, temos uma comparação de um para um (*1:1*).

- **Monitoramento:** (O escritor está sendo procurado?). Aplicações de monitoramento servem para determinar se uma pessoa encontra-se na lista de pessoas procuradas. Aplicações deste tipo podem ser vistas em aeroportos, fiscalizações e eventos públicos. Contudo, podemos utilizar esta técnica para monitorar pessoas através de sua assinatura ou manuscrito. Envolve uma comparação *1:N*, contudo, $N$ neste caso, é uma lista restrita.

Neste trabalho, o objetivo é verificar e identificar escritores através de documentos manuscritos, diferente do processo de reconhecimento de autoria que trabalha com o propósito de identificar autoria, não escritores. O processo de identificação e verificação de escritor trabalha unicamente com características do manuscrito, com o objetivo de saber quem escreveu aquele documento em papel. No processo de verificação e identificação de escritores, o conteúdo semântico do texto não é levado em consideração, diferente do

processo de reconhecimento de autoria, no qual tal característica é essencial para o bom desempenho do sistema [71].

De acordo com o método de aquisição da amostra manuscrita, o processo pode ser classificado como *on-line*, ou *off-line*. Para a abordagem *on-line* necessita-se de um hardware especial, tais como mesa digitalizadora, caneta sensível à pressão ou *tablet*. Já na abordagem *off-line*, o conteúdo encontra-se em papel (carta, contrato) sendo, posteriormente, digitalizado. Neste trabalho, utilizou-se a abordagem *off-line* devido a documentos manuscritos em papel serem utilizados como provas legais para questões judiciais.

Podemos classificar o texto em: dependente ou independente do conteúdo textual. Texto-dependente requer que todos os escritores escrevam o mesmo texto. Em texto-independente, o escritor não necessita escrever um texto padrão, escreve um texto próprio, não há um número mínimo ou máximo de palavras ou linhas (é comum existir um tema ao qual o escritor discorre sobre o mesmo). A utilização de técnicas alográficas em texto-dependente proporciona melhores taxas no processo de identificação, pois é possível a comparação de letra a letra ou palavra a palavra. Na abordagem textural, texto-dependente também tende a apresentar melhor desempenho, pois existe um número fixo de palavras. Um dos grandes problemas do texto-independente é o fato de que alguns escritores contribuem com pouquíssimo texto, tornando-se difícil a tarefa de identificação de escritores, pois não há quantidade de texto suficiente para representar um escritor.

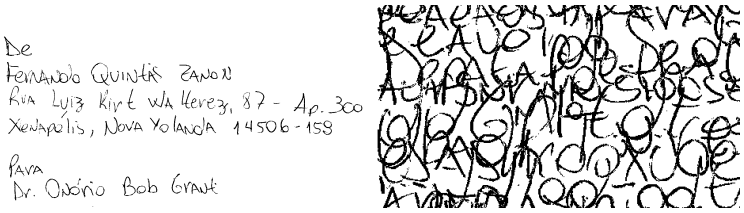
Segundo Justino et al. [50], a grafoscopia foi concebida com o intuito de esclarecer questões judiciais, tratando-se de uma área, cuja finalidade é a verificação da autenticidade de documentos a partir de características gráficas utilizadas na escrita deste documento. Cool et al. [24] e Justino [50] apresentam uma série de características individuais empregadas por peritos para análise da individualização da escrita, como: nível de habilidade, inclinação axial, forma caligráfica, descontinuidades, proporções, mínimos gráficos, pressão, alinhamento entre outras.

Baranoski [6] descreve um método comumente usado por peritos forenses, na qual o perito utiliza um conjunto com $n$ amostras de manuscritos com escrita desconhecida (referências - $R$) em comparação com a amostra de escrita questionada ($S$). O perito observa as diferenças entre as $L$ características grafoscópicas do conjunto de referência e da amostra questionada. Após este procedimento, toma-se uma decisão parcial. O laudo pericial resultante $D$ depende da soma dos resultados parciais obtidos das comparações dos pares (Documento de Referência / Documento Questionado).

Diversos trabalhos utilizam uma abordagem similar à empregada por peritos forenses para verificação e identificação de escritores [94, 92, 45]. Nestes casos, características alográficas são de suma importância para a verificação e identificação de escritor, tanto no contexto da perícia grafoscópica convencional, quanto nas abordagens computacionais, pois, através da frequência de ocorrências destas características é possível distinguir um escritor de outros.

Entretanto, a abordagem empregada baseia-se no conteúdo textural da imagem. As-

sim, utilizaremos técnicas de geração de textura proposta por Hanusiak [42], com intuito de gerar uma textura mais densa a partir do documento original. Conforme ilustrado na Figura 1.2(a), a ideia é retirar todo o espaço em branco existente entre linhas e palavras presentes no documento manuscrito. Desta forma, representaremos o escritor não por sua escrita propriamente dita, mas por uma textura gerada a partir da sua escrita. Logo, poderemos utilizar métodos que descrevem texturas para o processo de identificação e verificação de escritores. Para realização deste trabalho, utilizaremos documentos (cartas) digitalizados de diversos escritores em três diferentes línguas: Português, Inglês e Holandês.




(a) (b)

Figura 1.2: Figura (a): exemplo de carta manuscrita. Figura (b) textura gerada a partir de carta original.

Embasados nos trabalhos de [22, 72, 70] percebe-se que, por meio da dissimilaridade, podemos transformar um problema de $n$ classes em problemas com apenas duas classes. Em conjunto com a dissimilaridade, empregaremos conceitos da abordagem escritor-independente. A abordagem escritor-independente [70] possui duas vantagens em relação à abordagem escritor-dependente. A primeira é a independência do modelo, quanto às classes empregadas nos conjuntos de treinamento e teste, ou seja, não há necessidade que escritores, no conjunto de testes, possuam amostras no conjunto de treinamento. Em consequência disto, podem-se inserir novas classes no conjunto de teste, sem haver necessidade de retreinar um modelo. Desta forma, podem-se gerar modelos robustos, mesmo possuindo poucos exemplos genuínos de um mesmo escritor.

## 1.2 Motivação

A Ciência Forense, por meio de peritos, analisa a possibilidade da existência de fraudes e falsificações em documentos manuscritos; sendo, na grande maioria das vezes, referente a questões judiciais. Manuscritos podem ser também características-chave na resolução de crimes, na investigação de atentados terroristas, cartas de falso suicídio entre outros.

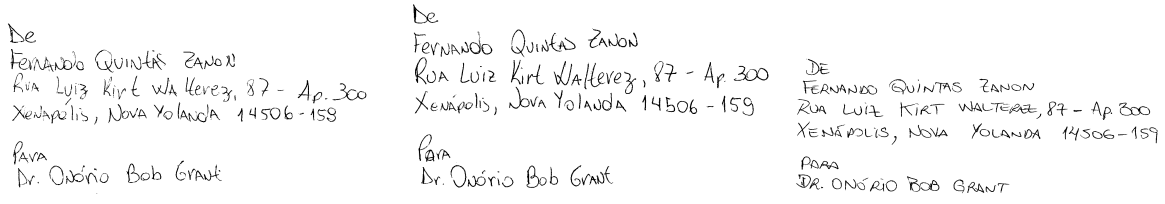
Textos manuscritos utilizados no intuito de verificar ou identificar quem escreveu determinado documento vêm sendo estudados há décadas [73], principalmente nos últimos anos devido aos grandes investimentos em segurança e ao crescente aumento da digitalização

de documentos. Assim, sistemas automáticos de verificação e identificação de escritores podem contribuir e auxiliar cientistas forenses na tomada de decisões, diminuindo a necessidade de mão de obra humana, o tempo de resposta e a imprecisão decorrente da subjetividade dos peritos devido à aplicação de técnicas grafométricas. Consequente a tudo isso, percebemos a necessidade de pesquisas na área visando construir um sistema automático ou semiautomático de alta precisão, independente de língua e que seja robusto, tanto para documentos com texto-dependente quanto para texto-independente.

## 1.3 Desafios

Um dos fatores, que leva sistemas de verificação e identificação de escritores não serem triviais, deve-se às fortes variações de características intrapessoais e a possíveis similaridades interpessoais. Outro problema é que a escrita de uma pessoa pode sofrer alterações ao longo dos anos, isso devido a uma série de fatores físicos e psicológicos intrínsecos a cada um. No processo de identificação de escritor, no qual temos uma comparação de $1 : N$, o nível de dificuldade aumenta consideravelmente.

Conforme Figura 1.3, pode-se observar variações de escrita de um mesmo escritor e também a similaridade entre a escrita de diferentes escritores pode ser alta. As Figuras 1.3(a), 1.3(b) e 1.3(c) demonstram tais propriedades.

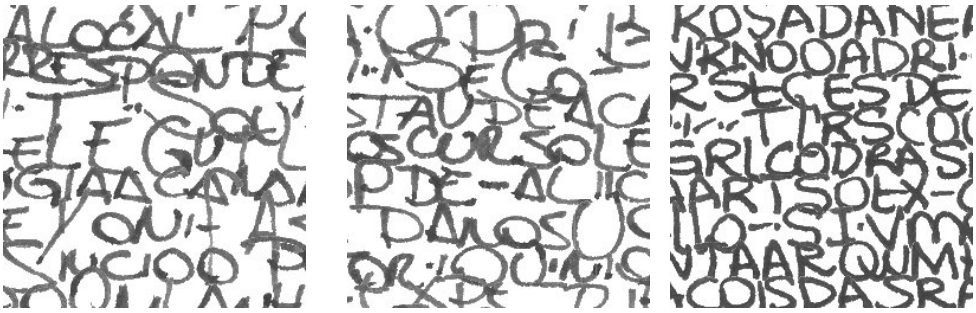


(a) (b) (c)

Figura 1.3: Variação intrapessoal existente entre escrita do mesmo escritor: (a) e (b). Similaridade existente entre amostras de dois diferentes escritores: (b) e (c).

Mesmo com a eliminação parcial dos espaços em branco, presentes nos documentos manuscritos, através do processo de geração do conteúdo textural, existe uma grande dificuldade no processo de verificação e identificação de escritores, devido às semelhanças interpessoais e variações intrapessoais. A Figura 1.4 apresenta tais semelhanças e diferenças. Percebe-se, então, que o primeiro grande desafio deste trabalho é a variabilidade entre classes.

Um dos grandes desafios em sistemas de identificação de escritores é o número de classes existentes. Conforme o número de escritores aumenta, a dificuldade devido a similaridades interclasses também aumenta, pois a complexidade em separar amostras de diferentes escritores que possuem similaridade na escrita é grande. Para superar este



(a) (b) (c)

Figura 1.4: Variação intrapessoal existente entre blocos de textura de um escritor: Figuras (a) e (b). Similaridade interpessoal entre blocos de textura de dois escritores: Figuras (b) e (c).

desafio, utilizamos conceitos de dissimilaridade, os quais transformam um problema de $n$ classes em duas classes.

Aplicações de verificação e identificação de escritores empregando características alográficas são eficientes de um modo geral. Entretanto, o sucesso do uso de características alográficas está relacionado ao desempenho da etapa de segmentação, etapa com alto custo computacional e com alto nível de complexidade e imprecisão. Assim, optamos por utilizar uma abordagem que não necessite do processo de segmentação, desta forma, propomos o uso descritores de textura para extrair características do documento. Percebendo que a densidade da textura poderia contribuir para um bom desempenho, empregamos técnicas de geração de textura apresentada por Hanusiak et al. [42].

Com isso, entende-se que o grande desafio deste trabalho está em combinar as etapas descritas, no intuito de construir um sistema de verificação e identificação de escritores robustos tanto para texto-dependente quanto para texto-independente, utilizando uma abordagem independente de escritor, a qual não necessita retreinar modelos ao inserir novas classes no conjunto de testes. Desta forma, o principal desafio é demonstrar que a técnica de escritor-independente não necessita de um grande número de classes (escritores) para gerar um modelo robusto, mas sim, de classes que, quando combinadas, possam gerar um modelo que apresente um bom desempenho para um sistema automático de verificação e identificação de escritor.

## 1.4 Objetivos

O principal objetivo deste trabalho é construir um sistema de identificação e verificação de escritor baseado nos conceitos de dissimilaridade e descritores de textura com intuito de ser robusto e independente do estilo da escrita e também da língua escrita.

Para atingir esse objetivo principal, destacam-se os seguintes objetivos marginais:

- Avaliar o impacto de diferentes descritores de textura para esta aplicação;
- Avaliar o desempenho do sistema, ao possuir diferentes números de referências para os conjuntos de treinamento e testes e o quão importante é o número de escritores no conjunto de treinamento;
- Avaliar os modelos gerados através da abordagem de seleção de escritores, observando o desempenho do mesmo;
- Verificar se, através da abordagem de seleção de escritores, é possível reduzir o número de escritores no conjunto de treinamento, contribuindo para uma melhora no desempenho no sistema;
- Avaliar o desempenho do sistema para bases isoladas de diferentes línguas com texto-dependente, texto-independente, estilo caixa alta e falsificação. Aplicar a abordagem proposta em uma grande base de dados formada através da união de bases em diferentes línguas.

A originalidade deste trabalho encontra-se fundamentada no uso de textura, em conjunto com a abordagem escritor-independente, aliada a um processo de seleção de escritores, no intuito de gerar modelos robustos, utilizando poucos escritores.

## 1.5 Contribuições

Os estudos realizados neste trabalho se destacam por apresentar uma abordagem inovadora para aplicações de identificação e verificação de escritores em documentos questionados, empregando poucos escritores no conjunto de treinamento.

Pode-se destacar ainda, as seguintes contribuições:

- Resultados a partir da classificação baseada na abordagem de dissimilaridade [12];
- A viabilidade do uso da abordagem independente de escritor comparada com a abordagem escritor-dependente [12];
- A avaliação de diferentes descritores de textura empregados em textura gerada através da escrita [61, 12];
- Uma abordagem robusta indiferente do estilo de escrita como texto-dependente, texto-independente e caixa alta [13];
- Um método eficiente para identificar falsificações [13];
- A proposta de seleção de escritores, deixando em aberto para ser testada em outras aplicações;

- Uma proposta robusta independente da língua e da quantidade de escritores [12, 13];
- Uma abordagem inovadora, a qual reduz o número de escritores no conjunto de treinamento, gerando modelos robustos, tal como empregando conjuntos até oito vezes maiores.

## 1.6 Organização

Este trabalho desenvolve-se ao longo de seis capítulos. Este capítulo contém uma breve descrição sobre o processo de verificação e identificação de escritores e a apresentação de alguns conceitos sobre o tema. No Capítulo 2 é apresentado um estudo sobre procedimentos utilizados, contribuindo com o leitor para um maior entendimento sobre técnicas e métodos computacionais a serem empregados nesta pesquisa. O capítulo 3 traz uma visão geral sobre o estado da arte. O Capítulo 4 mostra, em detalhes, o método proposto para a elaboração desta tese. No Capítulo 5 são apresentados os resultados obtidos através dos experimentos realizados ao longo desta pesquisa. No Capítulo 6 apontam-se as conclusões alcançadas por meio desta pesquisa.

# CAPÍTULO 2

# FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentadas algumas abordagens que norteiam esta tese. Temos como objetivo, dar suporte ao leitor, contribuindo para um melhor entendimento quanto aos métodos empregados neste trabalho. Entretanto, maior riqueza de detalhes poderá ser encontrada nas referências bibliográficas, aqui citadas. A seção 2.1 apresenta os descritores de textura que serão utilizados. Na seção 2.2 apontamos conceitos de dissimilaridade empregados neste trabalho. Na seção 2.3 são descritas as regras de combinação de classificadores a serem empregadas para combinar as saídas dos classificadores. A seção 2.4 finaliza o capítulo com alguns comentários.

## 2.1 Textura

Assim como a forma e a cor, a textura é facilmente percebida por um observador humano. Entretanto, mesmo a textura sendo um fenômeno de fácil reconhecimento e entendimento, é algo difícil de conceituar. Gonzalez e Woods [38] definem textura como sendo um conjunto de características estatísticas ou outras propriedades locais da imagem, que sejam constantes, com pouca variação ou aproximadamente periódicas.

A textura é uma importante característica para a análise de imagens em diferentes aplicações. Atualmente, o estudo de técnicas de análise e classificação de textura aborda áreas como: medicina, biometria, segurança, análise de imagens de satélites, entre tantas outras [43]. Desta maneira, métodos para descrever textura vêm sendo estudados há décadas [43, 64, 63, 21].

Rao et al. [76] apresentam um estudo sobre a percepção humana em relação à textura. Os autores, concluíram que existe um conjunto de características perceptuais, como contraste, repetitividade, granularidade, entre outros, que capturam diferentes aspectos das texturas. Tais aspectos podem ser observados na Figura 2.1.

Percebe-se, então, que a textura não pode ser definida através de um único pixel, mas sim, de um conjunto de pixels. Este conjunto pode ou não descrever um padrão de primitivas existente na textura. Estas primitivas são descritas como *TEXture ELement* (*Textel*). É comum existir uma variância significativa entre padrões de textura. Entretanto, a relação entre *Textels* deve ser suficientemente boa para diferenciar várias texturas.

Segundo Gonzalez e Woods [38], as principais abordagens para extrair descritores de textura são: Estatística, Estrutural e Espectral. Atualmente, a literatura apresenta diversas técnicas de extração de características de textura, [43, 64, 27, 40]. Neste trabalho, concentraremos nossos estudos nas abordagens estatística e estrutural, e isto se deve a

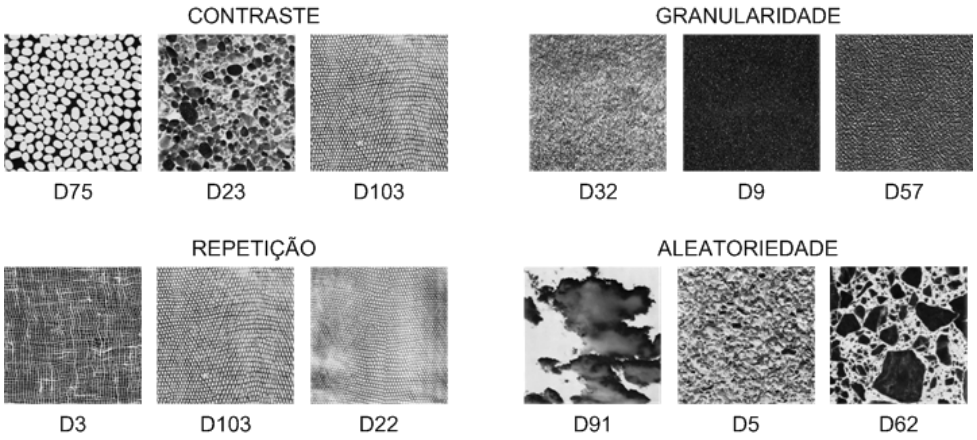



Figura 2.1: Exemplos de texturas selecionadas na base Brodatz [19].

três fatores: (*i*) O descritor estatístico GLCM irá funcionar como base de comparação para novos descritores; (*ii*) Os descritores estruturais LBP e LPQ vêm apresentando bons desempenhos nos mais variados estilos de textura, demonstrando ser, realmente, robustos [30]; (*iii*) Experimentos preliminares empregando técnicas de abordagem espectral, como Filtros de Gabor, apresentaram resultados inferiores ou próximos ao GLCM. Realizamos alguns experimentos com outros descritores estruturais como SIFT (*Scale Invariant Feature Transform*) [57] e SURF (*Speed-Up Robust Feature*) [8], entretanto, ambos apresentaram desempenhos inferiores aos alcançados com LBP e LPQ. Descreveremos a seguir os métodos estatísticos e estruturais.

### 2.1.1 Abordagem Estatística

Modelos estatísticos para descrição de textura têm por objetivo extrair medidas estatísticas de imagens digitais. Tais técnicas podem realizar uma contagem da ocorrência de níveis de cinza relativos aos pixels da imagem. Outra maneira bastante conhecida é verificar o modo como pixels com diferentes intensidades se relacionam na imagem. A seguir, descreveremos um método proposto por Haralick [43], de um descritor embasado em características estatísticas.

#### 2.1.1.1 GLCM

*Gray Level Co-ocurrence Matrix* (GLCM) é um método estatístico de segunda ordem, proposto por Haralick [43], em 1973, para descrever textura. O GLCM consegue descrever atributos existentes em imagens, como: suavidade, rugosidade, aspereza, granularidade, entre outros. A proposta de extrair descritores a partir de padrões estatísticos existentes na textura é relativamente simples e funcional.

A partir de uma imagem $I(x, y)$, uma matriz quadrada $n \times n$ é extraída da imagem, em que $n$ representa o número de níveis de cinza presente na imagem. O método GLCM consiste em encontrar medidas estatísticas a partir de uma matriz extraída da imagem.

Dois parâmetros estão diretamente relacionados ao método, sendo, a distância entre os pixels e o ângulo $\theta$, isto, a partir do pixel de interesse em relação a seus vizinhos. Para a obtenção das matrizes de co-ocorrência, considera-se a variação da distância e a direção entre pixels vizinhos, ou seja, de acordo com a orientação espacial, é calculada a probabilidade de existência de uma dada diferença entre dois pixels. Assim, uma matriz é extraída para cada direção e ângulo. Em geral, utilizam-se distâncias entre 1 e 5, e os ângulos mais comumente avaliados são: 0°, 45°, 90° e 135°.

A partir da matriz GLCM são extraídas algumas medidas, definidas como características. Quatorze medidas são descritas por Haralick [43]; as principais são:

- **Energia:** Pode ser encontrada como Segundo Momento Angular, basicamente avalia a uniformidade da textura em uma imagem. Sua fórmula é descrita por:

$$Energia = \sqrt{\sum_{i=0}^{N_g}\sum_{i=0}^{N_g}\{P(i,j)\}^2}. \tag{2.1}$$

- **Contraste:** Mede a presença de uma transição abrupta de níveis de cinza na imagem. Seu cálculo é dado por:

$$Contraste = \sum_{i=0}^{N_g}\sum_{i=0}^{N_g}|i-j|^2 P(i,j) \tag{2.2}$$

- **Homogeneidade:** Mede a regularidade presente na imagem. Sua fórmula é:

$$Homogeneidade = \sum_{i=0}^{N_g}\sum_{j=0}^{N_g}\frac{P(i,j)}{1+|i-j|^2}. \tag{2.3}$$

- **Correlação:** Mede a dependência no nível de cinza do pixel de interesse em relação a seus vizinhos. Sua fórmula é:

$$Correlação = \sum_{i=0}^{N_g}\sum_{j=0}^{N_g}\frac{(i-\mu_i)(j-\mu_j)}{\sigma_i^2\sigma_j^2}P(i,j) \tag{2.4}$$

Para as quatro Equações descritas anteriormente, $P(i,j)$ representa o conteúdo da coordenada $(i,j)$ da GLCM normalizada; $N_g$ representa o número dos diferentes níveis de cinza da imagem. Na Equação 2.4, $\sigma_i^2\sigma_j^2$ representam a variância de $i$ e $j$.

### 2.1.2 Abordagem Estrutural

Uma característica desta abordagem é proporcionar uma boa representação de textura para imagens com certa uniformidade. Entretanto, Puig et al. [75], avaliaram algumas

técnicas utilizando abordagem estrutural em diferentes espécies de textura, demonstrando que algumas técnicas estruturais funcionam muito bem, mesmo quando não existe uma boa homogeneidade entre amostras de textura. Métodos estruturais descrevem a textura a partir da relação espacial existente entre regiões ou primitivas presentes na imagem. Desta forma, é comum, na literatura, rotular estas primitivas de textura como *Textel*. Assim, a proposta de métodos embasados nesta abordagem é buscar a disposição existente entre regiões da imagem.

A seguir, descrevemos as duas abordagens utilizadas. A justificativa de empregarmos LBP e LPQ em nossos experimentos se deve aos excelentes resultados apresentados em literatura, [30], tanto para aplicações de identificação e verificação de escritores quanto para as mais diversas áreas [26, 62, 1].

#### 2.1.2.1 LBP

Método proposto por Ojala et al. [68] como medida complementar para o contraste local da imagem. Posteriormente, o método foi adaptado, tornando-se uma abordagem estrutural invariante à rotação para descrição de textura. *Local Binary Pattern* (LBP), baseia-se na conjuntura que padrões binários locais e a região de vizinhança de um pixel são características fundamentais na textura da imagem. Desta forma, o histograma formado a partir destas características é uma boa maneira para se representar a textura, ou seja, um ótimo descritor de textura.

A proposta inicial do LBP considera que cada pixel existente na imagem será, em algum momento, o ponto central para uma matriz de convolução de tamanho $3 \times 3$. Avaliando sua vizinhança em relação ao pixel central e considerando uma distância de um pixel, totalizam-se oito vizinhos. Ao comparar a intensidade do pixel central em relação a seus vizinhos, (caso seu vizinho possua intensidade maior que a do pixel central), este será classificado como 1, caso contrário 0. Logo após, é realizada uma multiplicação entre os valores binários e os decimais gerados, atribuindo ao pixel central o somatório desta função. A Figura 2.2 ilustra estas etapas.

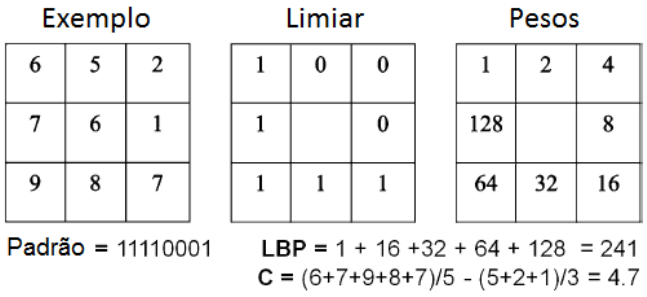



Figura 2.2: Modelo original do LBP. Adaptado de Maempa [64].

A partir da proposta inicial do LBP, Ojala et al. [68], associam a um pixel central um conjunto de amostras $P$, estando estas uniformemente espaçadas e distribuídas sobre

determinada circunferência de raio $R$, sendo seu centro o pixel central. Existem, agora, dois parâmetros que possuem forte relacionamento com o LBP, o $P$ que se refere ao número de vizinhos e o $R$ que remete à dimensão do raio. A Figura 2.3 demonstra alguns valores para $P$ e $R$.

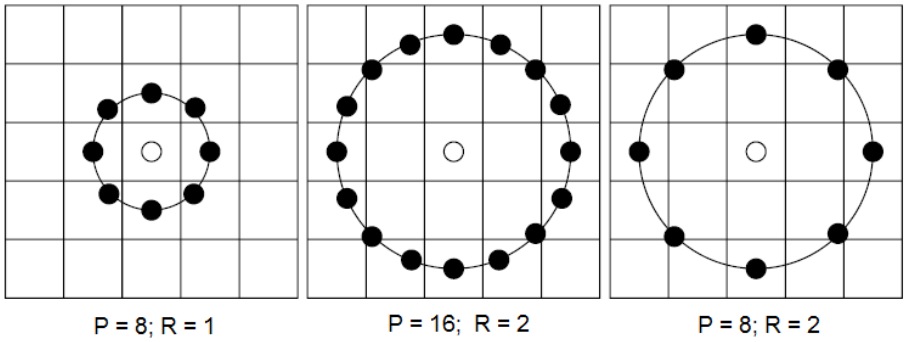



Figura 2.3: Diferentes valores de $P$ e $R$ para LBP.

Ao descrever a versão do LBP, os autores utilizam um rótulo (neste trabalho "u2"). Este rótulo, acompanhado do valor do raio $R$ e do número de vizinhos $P$, faz com que a definição do LBP seja, $LBP^{rotulo}_{P,R}$.

A partir das etapas anteriores, gera-se um histograma considerando um conjunto de padrões de transição de *bits*. Este padrão de transição é a ocorrência de, no máximo, duas transições do *bit* zero para um e vice-versa. Considerando oito vizinhos ($P = 8$), um total de 256 padrões de transições é gerado, dos quais apenas 58 atendem à definição de uniformidade descrita por Ojala et al. [68]. Os 198 padrões de transições restantes são contabilizados todos juntos, o que leva o descritor LBP possuir um total de 59 características. Atualmente, diversos trabalhos com resultados promissores são apresentados utilizando LBP, para os mais diferentes fins [97, 90, 75, 30, 53].

#### 2.1.2.2 LPQ

Embasado na abordagem proposta por Ojala et al. [68] para um descritor de texturas local, Ojansivu e Heikkilä [69] apresentam um método local para descrição de textura, descrevendo-o como relacionado e complementar ao LBP: o LPQ (*Local Phase Quantization*). Sua principal característica é a robustez para imagens borradas ou afetadas por iluminação não uniforme. Os experimentos descritos pelos autores mostram ótimas taxas de desempenho do método para imagens borradas, quando comparadas com outros métodos como LBP ou Filtros de Gabor.

A proposta deste descritor é que, a partir de cada pixel $x$ de uma imagem $N$, possamos representar a textura dos pixels, considerando uma vizinhança retangular $V_x$ ($m \times m$). O método tem sua base nas propriedades de espectro de fases de uma Transformada de Fourier de Curto Termo (STFT). A STFT $\hat{f}_{u_i}(x)$ para a imagem $f(x)$ é dada pela Equação 2.5.

$$\hat{f}_{u_i}(x) = (f * \phi_{u_i})(x) \tag{2.5}$$

sendo o filtro $\phi_{u_i}$ ($m \times m$) dado pela equação 2.6.

$$\phi_{u_i} = \ e^{-j2\pi u_i^T y} \mid y \in \mathbb{Z}^2 \ \|y\| \infty \leq r \tag{2.6}$$

onde $*$ denota a convolução de $\phi_{u_i}$ em $f$, $j = \sqrt{-1}$, $r = (m-1)/2$ e o vetor base da 2D-DFT $u_i$ na frequência $i$.

O LPQ leva em conta apenas quatro coeficientes complexos dos componentes real e imaginários da Transformada de Fourier, especificamente $u_1 = [a, 0]^T$, $u_2 = [0, a]^T$, $u_3 = [a, a]^T$ e $u_4 = [a, -a]^T$, sendo $a = 1/m$. Desta forma, são computados pela Equação 2.7, na qual $f(x)$ constitui um vetor através dos $m^2$ pixels da vizinhança $V_x$. Assim, é aplicado este procedimento para cada pixel da imagem, de forma que $\Phi_{u_i}$ representa o vetor base da 2D-DFT na frequência $i$ ($\phi_{u_i}$).

$$\hat{f}_{u_i}(x) = \Phi_{u_i}^T f(x) \tag{2.7}$$

Embasado na propriedade que a fase e não a magnitude possui a maior parte das informações da Transformada de Fourier, um processo de redução de dimensionalidade com relação aos oito valores complexos obtidos é realizado. Sendo $F = [f(x_1), f(x_2), ..., f(x_{n^2})]$ e a matriz $m^2 \times N^2$, a qual compreende a vizinhança de todos os pixels na imagem e $\Phi = [\Phi_R, \Phi_I]^T$, em que $\Phi_R = Re[\Phi_{u_1}, \Phi_{u_2}, \Phi_{u_3}, \Phi_{u_4}]$ representa sua parte real e $\Phi_I = Im[\Phi_{u_1}, \Phi_{u_2}, \Phi_{u_3}, \Phi_{u_4}]$ sua parte imaginária. Temos, então, que a matriz de transformação STFT é dada por $\hat{F} = wF$, onde $\hat{F}$ é uma matriz de valores reais e dimensão $8 \times N^2$.

Por fim, as etapas de decorrelação e quantização são aplicadas. O objetivo desta decorrelação, antes mesmo do processo de quantização, visa maximizar a informação de textura retida referente aos sinais das componentes real e imaginária. Ojansivu e Heikkilä [69] assumem que a função $f(x)$ de uma imagem é resultado de um processo de primeira ordem de Markov, onde o coeficiente de correlação entre pixels adjacentes $x_i$ e $x_j$ está relacionado exponencialmente com sua distância $d_{ij}^s$. Para o vetor $\vec{f}$ é definida uma matriz de covariância $C(m^2 \times m^2)$ dada pela Equação 2.8, na qual a matriz de covariância dos coeficientes de Fourier pode ser obtida por $D = \Phi C \Phi^T$. Tais coeficientes podem deixar de ser correlacionados através da transformação $E = V^T \hat{F}$, na qual $V$ é uma matriz ortogonal ($8 \times 8$) derivada dos valores singulares de decomposição (*Singular Value Decomposition*- SVD) de uma matriz $D$, que é $D' = V^T D V$.

$$C_{i,j} = \sigma^{||x_i - x_j||} \tag{2.8}$$

Os coeficientes quantizados através da Equação 2.9, na qual $e_{i,j} \in E$ são representados por valores inteiros entre 0 e 255, através de codificação binária pela Equação 2.10. O

vetor de características gerado através do descritor LPQ é dado por um histograma com 256 posições dos valores inteiros, computados para todas as posições da imagem.

$$q_{i,j} = \begin{cases} 1, & \text{se } e_{i,j} \geqslant 0 \\ 0, & \text{se } e_{i,j} < 0 \end{cases} \tag{2.9}$$

$$b_j = \sum_{i=0}^{7} q_{i,j} 2^i \tag{2.10}$$

Devido a algumas propriedades básicas da Transformada de Fourier em conjunto com o modelo proposto, o LPQ pode ser descrito como um descritor invariante a borramentos simétricos, o que inclui movimentação e falta de foco. Através das informações, considerando a fase, o método também apresenta invariância a mudanças uniformes de iluminação. Ojansivu e Heikkilä em [69] descrevem com maiores detalhes o descritor LPQ.

## 2.2 Representação da Dissimilaridade

No processo de classificar padrões, um observador humano pode utilizar-se de diferentes métodos para rotular algo. Contudo, qual abordagem utilizada é mais eficiente para classificação? Qual a primeira etapa para categorização do processo? Seria a forma, cor ou peso características discriminantes ? Ou ainda, a percepção que alguns objetos são de algum modo mais similar e outros possuírem maiores diferenças?

Com base na ideia de semelhanças e diferenças, Pekalska e Duin [72] abordam aspectos da representação por dissimilaridade, de forma que a teoria utilizada nesta abordagem é similar à realizada por humanos. A observação das diferenças e semelhanças entre objetos parece ser uma atitude bastante simplória, entretanto, é utilizada constantemente com o intuito de rotular elementos desconhecidos, como animais, plantas entre outros. Considerando tal abordagem, estudos visando representar um modelo, utilizando as diferenças entre objetos, apresentam-se como uma proposta interessante em diferentes áreas [23, 70, 80]. Recentemente Eskander et al. e Rivard et al. apresentaram vários trabalhos usando dissimilaridade aplicados à verificação de assinatura *off-line* [33, 78, 31, 32]. Comumente encontramos trabalhos cuja proposta de representação da dissimilaridade é apresentada como transformação dicotômica [77].

Métodos de representação da dissimilaridade possuem vantagens em relação a outros métodos. A primeira vantagem é a transposição de um aparente problema de reconhecimento de padrões, no qual o número de classes pode ser alta, ou não específica, em um problema de duas classes [23]. A outra é a não necessidade de retreinar um modelo quando inseridas novas classes no conjunto de teste.

Cha e Shihari [23] descrevem a transformação de um problema de múltiplas classes para um problema binário. Pekalska e Duin [72] introduzem a ideia de representar as relações entre objetos através de dissimilaridade, a que eles chamam Representação da Dissimila-

ridade. Este conceito descreve cada objeto $x$ através de suas diferenças a um conjunto de objetos de protótipo, definido como conjunto de representação $R$. Assim, cada objeto $x$ é representado por um vetor de dissemelhanças $D(x, R) = [d(x, r_1), d(x, r_2), ..., d(x, r_n)]$ para os $r_j \in R$ objetos.

Seja $R$ um conjunto de representação composta de $n$ objetos. Um conjunto de treinamento $T$ com $m$ objetos é representado por $D(T, R)$ por uma matriz de dissimilaridade $m$. Através da abordagem de representação da dissimilaridade, a classificação de um novo objeto $x$ representado por $D(x, R)$ é realizada usando a regra do vizinho mais próximo. Ao objeto $x$ é atribuído a classe de seu vizinho mais próximo, sendo a classe de representação do objeto $r_j$ dada por $d(x, r_j) = min_{r \in R} D(x, R)$. O ponto chave, aqui, é que as diferenças devem ser pequenas para objetos semelhantes (pertencentes à mesma classe) e, grande para objetos distintos.

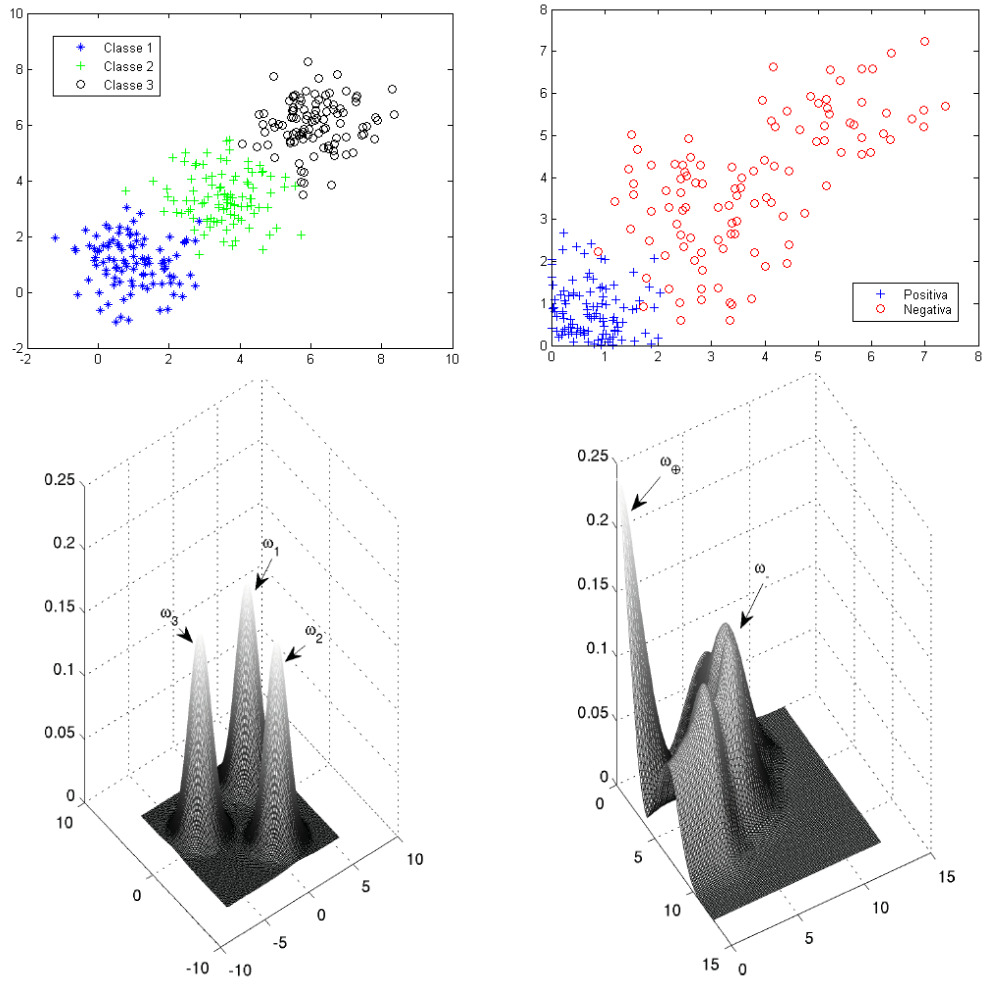
Através das distâncias intraclasses e interclasses é possível representar qualquer problema com um número finito de classes em apenas duas classes, utilizando a abordagem de dissimilaridade. Neste trabalho, os escritores caracterizam as classes e os documentos por eles escritos, que constituem as amostras. As características utilizadas, neste trabalho, referem-se a atributos de diferentes descritores de textura. Pekalska e Duin [72] descrevem que a abordagem pode ser viável quando a descrição baseada em características de objetos forem de difícil obtenção ou se apresentarem ineficientes para fins de aprendizagem.

Assim, nossa proposta é extrair vetores de características usando descritores de textura de ambos os conjuntos de escritores $(R)$ e $(S)$ e, posteriormente, computar os vetores de dissimilaridade. A partir de duas amostras da mesma classe, distâncias intraclasse são primeiramente computadas. A ideia é que a diferença entre amostras da mesma classe gerem vetores com componentes próximos de 0, caso contrário (interclasses), os componentes devem estar longe de 0. Isso é totalmente verdadeiro em condições favoráveis. Entretanto, não é possível assegurar, pois algumas classes, ao serem descritas em um novo espaço, devido a diversos fatores, como alta variação entre amostras do mesmo escritor ou ainda uma alta similaridade entre amostras de diferentes escritores, podem gerar uma maior confusão na transposição de domínios.

A Figura 2.4 apresenta a transformação de um problema de três para duas classes. Fica claro, em algumas instâncias da Figura 2.4(a) e após a transposição de espaço (Figura 2.4(b)), a proximidade entre amostras da mesma classe com amostras de diferentes classes.

A Figura 2.4(a) representa três diferentes classes, ($\omega_1$, $\omega_2$, $\omega_3$), de forma que cada uma é representada por um vetor bidimensional e tridimensional. A transformação deste problema de um espaço de características para um espaço de dissimilaridade é realizada gerando vetores de distâncias entre amostras de uma mesma classe, caracterizando estas como distâncias intraclasse, denotada por $x_{(+)}$. Para geração das distâncias interclasses, temos um vetor de distâncias através de amostras de diferentes classes, sendo denotada por $x_{(-)}$ (Figura 2.4(b)).

Sistemas baseados na abordagem de dissimilaridade necessitam de conjuntos de da-




(a) (b)

Figura 2.4: Na Figura (a) representamos características de três diferentes classes. Na Figura (b), representamos a transformação de três para duas classes. Adaptado de [77].

dos para treinar um modelo de aprendizagem de máquina. Como descrito anteriormente, nesta abordagem, as classes (representadas pelos escritores) presentes no conjunto de treinamento não fazem parte do conjunto de teste. O Algoritmo 1 resume os procedimentos de treinamento.

**Algoritmo 1** Treinamento

1: **Entrada:** Classes que serão usadas para treinar o modelo de aprendizagem de máquina. Como afirmado anteriormente, uma vantagem dessa abordagem é que podemos utilizar conjuntos disjuntos para treinamento e testes.
2: **Saída:** Modelo de Aprendizado de Máquina $\Phi$, treinados para discriminar entre vetores de dissimilaridade positivos e negativos.
3: ExemplosPositivos $\leftarrow 0$
4: **for** $i \leftarrow 1$ to $NumeroClassesTreino$ **do**
5: Escolhe aleatoriamente um conjunto ($R$) de $n$ imagens para ser utilizado como referência
6: */* Extrai características texturais das imagens de referência */*
7: **for** $j \leftarrow 1$ to $n$ **do**
8: $V_{ij} \leftarrow$ Vetor de Características extraído de $R_j$
9: **end for**
10: */* Computa exemplos positivos de dissimilaridade. Diferenças entre vetores da mesma classe */*
11: **for** $j \leftarrow 1$ to $n$ **do**
12: **for** $k \leftarrow j + 1$ to $n$ **do**
13: $Z_{(+)} \leftarrow |V_{ij} - V_{ik}|$
14: ExemplosPositivos++
15: **end for**
16: **end for**
17: **end for**
18: */* Computa exemplos Negativos de dissimilaridade. Diferenças entre vetores de diferentes classes */*
19: ExemplosNegativos $\leftarrow 0$
20: **while** (ExemplosNegativos ¡ ExemplosPositivos) **do**
21: $a \leftarrow$ Vetor de Características Extraído Aleatoriamente a partir de $V$
22: $b \leftarrow$ Vetor de Características Extraído Aleatoriamente a partir de $V$, mas a partir de uma classe diferente de um $a$
23: $Z_{(-)} \leftarrow |a - b|$
24: ExemplosNegativos++
25: **end while**
$\Phi \leftarrow$ Conjunto de Treinamento para Aprendizagem de Máquina $(Z_{(+)}, Z_{(-)})$
26: **return** $\Phi$

Após o processo de treinamento dos classificadores de dissimilaridade, o teste é feito através do Algoritmo 2. Na linha 6, Algoritmo 2, várias funções podem ser utilizadas para combinar a decisão parcial do classificador. Em nossos experimentos, a função que proporcionou os melhores resultados foi a regra da soma [52].

**Algoritmo 2** Teste

1: **Entrada:** Vetor de Características dos padrões de Teste ($Q$), $k$ referências de todo o conjunto de Teste ($S$) empregado no sistema.
2: **Saída:** A classe a que $Q$ deve ser atribuído.
3: **for** $i \leftarrow 1$ to $NumeroClassesTeste$ **do**
4: $Z_k \leftarrow |Q - S_{ik}|$ */*Computa os vetores de dissimilaridade.*/*
5: Saídas $\leftarrow \Phi Z_k$. */*Classificação dos k vetores de dissimilaridade usando o classificador gerado anteriormente.*/*
6: DecisãoParcial $\leftarrow$ Combina(Saídas). */*Combina as k* ***Saídas*** *em uma única decisão parcial.*/*
7: **end for**
8: **return** $max$(DecisãoParcial) */*Retorna à classe que maximiza as decisões parciais. */*

## 2.3 Combinação de Classificadores

Esta seção apresenta uma síntese sobre esquemas de combinação de classificadores, embasados nos trabalhos de Kittler et al. [52], Jain et al. [48] e Dietterich [28].

Segundo Kittler et al. [52] diversos são os motivos para a combinação de classificadores, um deles é o fato do desempenho apresentado por alguns esquemas que superam consistentemente o classificador de melhor desempenho individual. Neste trabalho, utilizamos um esquema de dissimilaridade que processa vários fragmentos de textura de um mesmo documento. Assim, para chegarmos a uma decisão final, utilizamos esquemas de combinação para combinar as saídas dos $N$ fragmentos empregados no esquema de dissimilaridade.

Partindo do pressuposto que, possuindo uma decisão consensual em vez de uma decisão individual, podemos alcançar desempenhos melhores, a abordagem se apresenta bastante motivadora. Entretanto, para tal, dependemos de que todos os membros do agrupamento cometam erros de classificação independentes, pois, somente desta maneira, poderemos garantir que a combinação de suas decisões individuais poderá implicar a melhoria do desempenho de classificação. Jain et al. [48] apresenta a ideia do uso de Sistemas com Múltiplos Classificadores (MCSs), nos quais consideram-se que classificadores diferentes, produzindo desempenhos superiores aos demais em diferentes momentos, impossibilitam a escolha de um único classificador, tornando necessário o emprego dos MCSs. Jain et al. abordam também sobre a perspectiva dos diferentes tipos de saídas dos classificadores, classificando-os em: abstrato, *ranking* e probabilístico. Assim, se o classificador provê apenas o rótulo da classe predita como saída, este é classificado como abstrato. Se o classificador retorna uma lista ordenada com as $n$ possíveis classes candidatas, é denominado *ranking*, por fim, é comum o uso de classificadores que proveem, como saída, valores de probabilidades *a posteriori* (*scores* ou estimativas de probabilidades); estes, denominam-se probabilísticos. Neste trabalho, empregamos basicamente classificadores com saídas probabilísticas. Um aspecto interessante é a forma como se podem combinar as saídas de classificadores ao possuirmos estimativas de probabilidades. Os esquemas de

combinação apresentados, neste trabalho, são métodos que independem dos dados propostos por Kittler et al. [52], ou seja, não são influenciados por dados do treinamento, considerados esquemas de agregação simples, como exemplo: Voto Majoritário, Soma, Produto, Média, Mediana, Máximo e Mínimo.

De acordo com o trabalho de Kittler et al. [52], considera-se um problema de reconhecimento de padrões, no qual assumimos o padrão $Z$ para uma das $m$ possíveis classes $(\omega_1, ..., \omega_m)$. Supondo que existem $R$ classificadores, de forma que cada um representa uma classe por um vetor de características distintas e assumindo que o vetor usado pelo *i-ésimo* classificador é $x_i$. Na dimensão do espaço, cada classe $\omega_k$ é modelada por uma função densidade de probabilidade $p(x_i|\omega_k)$ sendo sua probabilidade *a priori* de ocorrência denotada por $P(\omega_k)$.

Conforme a teoria Bayesiana, dada a dimensão $x_i, i = 1, ..., R$, o padrão $Z$ deve ser atribuído à classe $\omega_j$, a qual oferece a probabilidade *a posteriori*, cuja a interpretação é máxima, ou seja:

$$\begin{aligned} &atribuir\ Z \rightarrow w_j\ se \\ P(w_j|x_1, ..., x_R) &= max_k P(w_k|x_1, ..., x_R) \end{aligned} \tag{2.11}$$

A regra de decisão de Bayes 2.11, estabelece que, para utilizar toda a informação existente para se chegar a uma decisão correta, é essencial calcular as probabilidades de várias hipóteses, considerando, simultaneamente, todas as medidas. Mesmo esta sendo uma declaração sobre a correta classificação, pode ser uma proposição não viável.

Para computar a probabilidade *a posteriori*, dependemos do conhecimento de medidas estatísticas de alta ordem, descritas em termos de funções de densidade de probabilidade conjunta $p(x_i, ..., x_R|\omega_k)$, que seria difícil para inferir. Tenta-se simplificar a regra 2.11 e exprimí-la em termos de apoio à decisão dos classificadores individuais, onde cada um explora somente as informações dadas pelo seu vetor de característica $x_i$. Sendo assim, consegue-se construir uma regra de decisão computacional mais eficiente, através de regras de combinação que são comumente utilizadas na prática. Desta forma, ao reescrevermos a probabilidade *a posteriori* $p(\omega_k|x_1, ..., x_R)$, utilizando o teorema de Bayes, teremos:

$$P(w_k|x_1, ..., x_R) = \frac{p(x_1, ..., x_R|w_k)P(w_k)}{p(x_1, ..., x_R)} \tag{2.12}$$

no qual $p(x_1, ..., x_R)$ é uma medida incondicional da densidade de probabilidade conjunta. É apresentada, então, uma medida de distribuição condicional na equação 2.13:

$$P(x_1, ..., x_R) = \sum_{j=1}^{m} p(x_1, ..., x_R|w_j)P(w_j) \tag{2.13}$$

Kittler et al. [52], após uma série de deduções, a partir da Equação 2.12, descrevem que, dado o conjunto de distribuições de probabilidade das medidas extraídas pelos classificadores, tem-se uma das regras de combinações a seguir.

### 2.3.1 Regra do Produto

A regra de decisão definida na Equação 2.14 quantifica a probabilidade de uma hipótese ser combinada com a probabilidade *a posteriori*, gerada por classificadores individuais, através da regra do produto. Conclui-se que esta é uma regra eficientemente severa.

$$p^{-(R-1)}w_j \prod_{i=1}^{R} P(w_j|x_i) = max_{k=1}^{m} P^{-(R-1)}(w_k) \prod_{i=1}^{R} P(w_k|x_i) \quad \begin{array}{l} \\ atribuir\ Z \rightarrow w_j\ se \end{array} \tag{2.14}$$

### 2.3.2 Regra da Soma

Para a regra da soma, consideraremos a regra do produto (2.14) em maiores detalhes. Em alguns casos, assume-se que a probabilidade *a posteriori* calculada pelo respectivo classificador não diferenciará drasticamente da probabilidade *a priori*. Esta hipótese pode ser satisfeita quando a disposição da informação for muito ambígua, devido ao alto nível de ruído. Nesta situação, podemos assumir que a probabilidade *a posteriori* pode ser expressa por meio da regra de decisão da soma:

$$(1-R)P(w_j) + \sum_{i=1}^{R} P(w_j|x_i) = max_{k=1}^{m} \left[ (1-R)P(w_k) + \sum_{i=1}^{R} P(w_k|x_i) \right] \quad \begin{array}{l} \\ atribuir\ Z \rightarrow w_j\ se \end{array} \tag{2.15}$$

Kittler [52] faz uma breve reflexão sobre as regras da soma e do produto, em que, talvez, o ponto mais importante seja o fato de todas as regras de decisões, a serem derivadas destas, serem largamente usadas na prática. Através dos esquemas de combinação apresentados por Kittler, a partir das regras de decisão do produto (2.14) e da soma (2.15), outras regras de decisão podem ser desenvolvidas.

### 2.3.3 Regra do Máximo

A partir da Equação 2.15, é possível aproximar a regra da soma pelas máximas probabilidades *a posteriori*, de forma que, assumindo probabilidades *a priori* iguais, obtém-se a regra do máximo, Equação 2.16:

$$atribuir\ Z \rightarrow w_j\ se$$
$$max_{i=1}^{R} P(w_j|x_i) = max_{k=1}^{m} max_{i=1}^{R} P(w_k|x_i) \tag{2.16}$$

### 2.3.4 Regra da Mediana

A regra da mediana atribui um padrão à classe, cuja probabilidade *a posteriori* seja máxima. Entretanto, se um dos classificadores de saída adotar uma probabilidade *a posteriori* com um desvio muito grande das demais classes, isto afetará a média, podendo conduzir a uma decisão incorreta. Sabendo disto, o mais adequado é basear a decisão de combinação na mediana da probabilidade *a posteriori*, levando para a seguinte regra:

$$atribuir\ Z \rightarrow w_j\ se$$
$$med_{i=1}^{R} P(w_j|x_i) = max_{k=1}^{m} med_{i=1}^{R} P(w_k|x_i) \tag{2.17}$$

### 2.3.5 Regra do Voto Majoritário

A partir da Equação 2.15, assumindo a probabilidade *a priori* e o enrijecimento das probabilidades, temos:

$$atribuir\ Z \rightarrow w_j\ se$$
$$\sum_{i=1}^{R} \Delta_{ji} = max_{k=1}^{m} \sum_{i=1}^{R} \Delta_{ki} \tag{2.18}$$

Com relação à soma do lado direito da equação (2.18), para cada $w_k$ temos a contagem dos votos recebidos para dada hipótese dos classificadores individuais. Assim, a classe com o maior número de votos é selecionada pela decisão da maioria.

## 2.4 Comentários

Em geral, procuramos descrever as principais técnicas a serem empregadas nesta tese. Contudo, algumas abordagens, bastante difundidas na comunidade acadêmica, não foram descritas neste trabalho, pois, centenas de trabalhos os descrevem em riqueza de detalhes. Um exemplo disso é o classificador SVM proposto por Vapnik [96] a ser empregado em nossos experimentos. Atualmente, há um acervo imenso de informações a respeito do mesmo. Empregaremos Algoritmos Genéticos como método de busca para realização deste trabalho, o mesmo também não foi descrito neste capítulo, pois, embasados nos trabalhos de John Holland [46], centenas de livros e artigos expõem, em detalhes, o método de busca baseado em seleção.

# CAPÍTULO 3

# ESTADO DA ARTE

Neste capítulo, serão descritos o estado da arte de sistemas de identificação e verificação de escritores, juntamente com as bases de dados a serem utilizadas neste trabalho. Sebastiani e Fabrizio [89] discutem sobre a dificuldade existente em comparar resultados obtidos de diversos trabalhos, utilizando diferentes bases de dados, as quais, muitas vezes, não são validadas pela comunidade acadêmica nem ao menos descritas. Desta forma, na seção 3.1 descreveremos algumas bases que vêm sendo empregadas em trabalhos recentes.

## 3.1 Bases de Dados

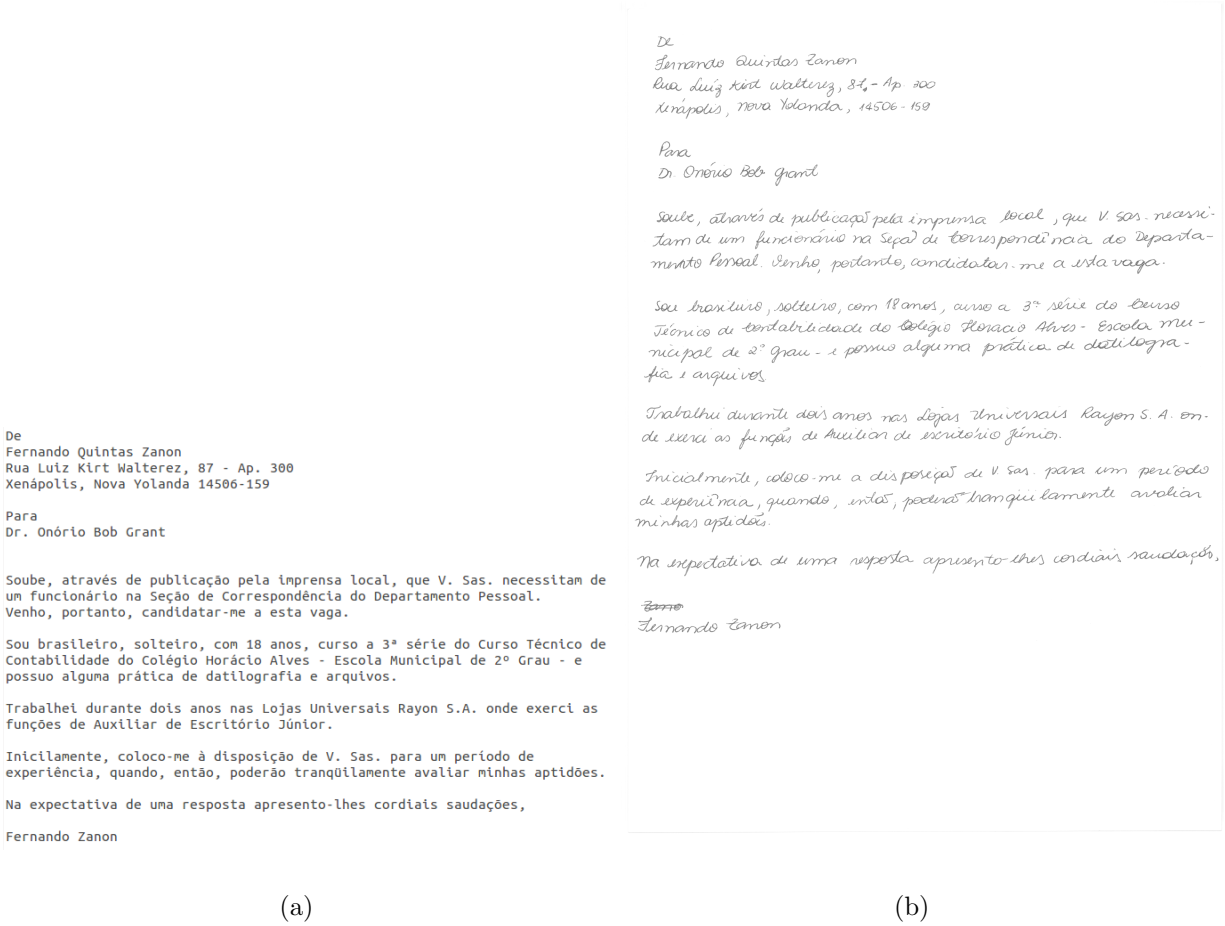
Atualmente, existem diversas bases de dados, como: CEDAR, NIST, CENPARMI, PSI, ETL9, PE92, RIMES, entre outras [59]. Entretanto, para este trabalho, optamos por empregar três bases públicas de diferentes línguas, sendo elas: BFL (*Brazilian Forensic Letter Database*), IAM (*Institut für Informatik und angewandte Mathematik*) e *Firemaker*. As bases também possuem características diferentes quanto ao tipo de texto, ou seja, texto-dependente e texto-independente. A base *Firemaker* trata-se de uma base mais completa, com a qual poderemos avaliar o desempenho em diferentes estilos de escritas, como: caixa-alta, texto-dependente, texto-independente e falsificação. A base IAM, além de ser muito empregada, possui um grande número de escritores e tem uma alta variação quanto ao número de cartas por escritor e o número de linhas escritas. Por fim, a base BFL é uma base de língua Portuguesa, com texto-dependente e um significativo número de escritores. Tais características foram decisivas no emprego destas três bases de dados para a realização deste trabalho. A seguir, descreveremos com detalhes as bases BFL, IAM e *Firemaker*.

### 3.1.1 Base BFL

Proposta por Baranoski [6], a *Brazilian Forensic Letter Database* (BFL) foi obtida de 2002 a 2005, através da colheita com pessoas voluntárias. Foi elaborada devido à falta de uma base escrita na língua Portuguesa. Todo o conteúdo da carta foi elaborado de forma a maximizar o conjunto de letras do alfabeto da língua Portuguesa (minúsculas e maiúsculas). Além de possuir um léxico de 124 palavras, conta também com algumas particularidades da língua, como símbolos de acentuação, tais como o til, cedilha, acento circunflexo, acento grave, acento agudo e pingo nos "is", além de mínimos gráficos. Segundo os autores, todas as regras quanto à transcrição foram obedecidas, como: uso de caneta esferográfica azul ou preta e a escrita do texto sem auxílio de linhas-guia. A base

BFL trata-se de uma base texto-dependente, na qual todos os escritores escrevem a punho o mesmo conteúdo textual. A Figura 3.1(a) apresenta o texto reproduzido pelos escritores.

As imagens digitalizadas possuem 300 *dpi*, 256 tons de cinza e estão salvas no formato *bitmap* (.bmp). Composta por 315 escritores, em que cada um concede três amostras, totalizando 945 cartas. A Figura 3.1(b), apresenta a imagem original de um manuscrito da base BFL.




(a) (b)

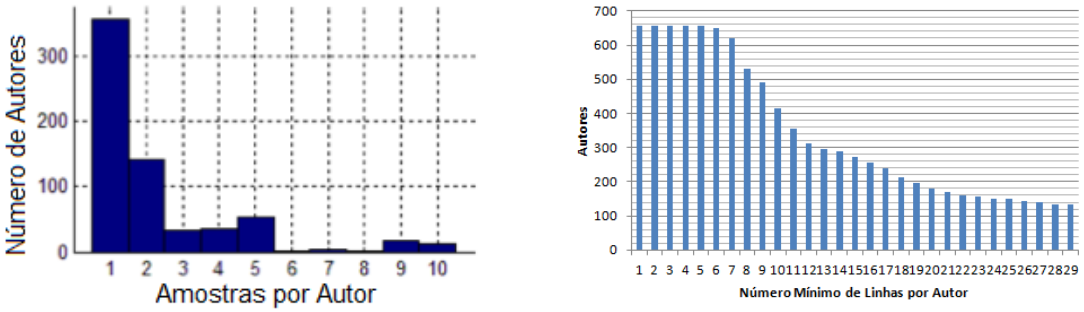
Figura 3.1: A Figura (a) apresenta o conteúdo textual da base BFL reproduzido pelos escritores. Na Figura (b), temos uma amostra redigida a próprio punho por um dos escritores.

### 3.1.2 Base IAM

A base IAM, encontra-se na língua Inglesa, sendo a quantidade de conteúdo escrito bastante variado, pois, trata-se de uma base texto-independente. A primeira versão da base IAM, de Outubro de 2002, é descrita por Marti e Bunke [59] e conta com 115.320 casos de palavras manuscritas, distribuídas em 13.353 linhas de texto. Produzida por cerca de 400 escritores, possui um léxico com 10.841 palavras diferentes. No entanto, o número de amostras manuscritas coletadas por escritor varia bastante (1 a 59 amostras por escritor),

sendo que a grande maioria (350) contribui com uma amostra. Na Figura 3.2(a), mostramos a distribuição de amostras por escritor. A Figura 3.2(b) apresenta a quantidade mínima de linhas escrita por escritor.

Atualmente, a base IAM 3.0 contém amostras de 657 escritores e encontra-se disponível online [1]. A Figura 3.3 apresenta alguns exemplares da base IAM.




(a) (b)

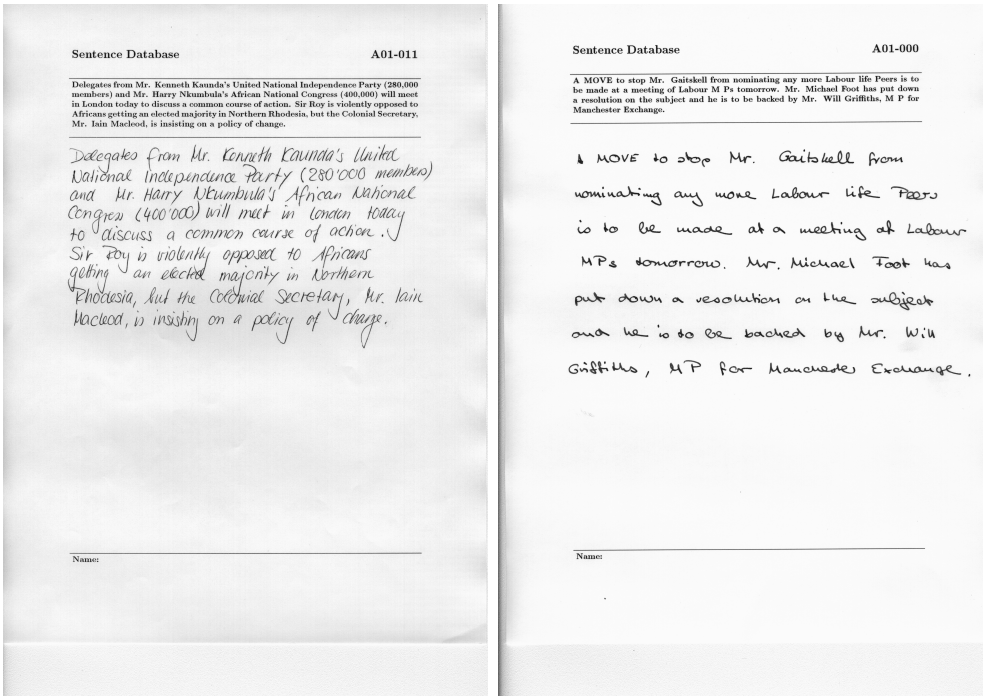
Figura 3.2: Na Figura (a) demonstramos a distribuição das amostras da base IAM. Na Figura (b), temos a distribuição de linhas por escritor.

Existem dois grandes desafios ao empregar a base IAM: o primeiro é o grande número de escritores. O segundo é a quantidade de texto manuscrito, cedido por cada escritor, pois sendo uma base de texto-independente, existe uma variação muito grande em relação ao número de palavras escritas, deixando o problema ainda mais desafiador. Do total de 657 escritores presentes na base IAM, sete deles foram excluídos, pois continham uma quantidade de texto manuscrito muito pequena, em que era impossível gerar blocos de textura com os mesmos. Brink et al. [16] analisa o impacto da quantidade de texto, necessário para sistemas de verificação e identificação de escritor. Entretanto, o autor faz uma análise quanto as características locais, descrevendo que o mínimo necessário seria algo entre 100 e 200 caracteres. Contudo, para a abordagem de geração de textura utilizada neste trabalho, constatamos que o ideal seria, no mínimo, sete linhas de texto. Assim, os escritores excluídos tinham uma ou duas linhas, no máximo.

### 3.1.3 Base *Firemaker*

A base *Firemaker* [88] criada em 2003, vem sendo utilizada em diversos trabalhos [88, 87, 18], apresentando-se como uma base bastante desafiadora, pois trabalhos recentes demonstram altas taxas de erros, quando comparadas a outras bases, como demonstra Brink et al. [18]. Devido a isto e ao fato desta base utilizar vários estilos de escrita, incluindo a falsificação, utilizaremos a base *Firemaker* neste trabalho.

[1] ¡*http://www.iam.unibe.ch/fki/databases/iam-handwriting-database*¿




(a) (b)

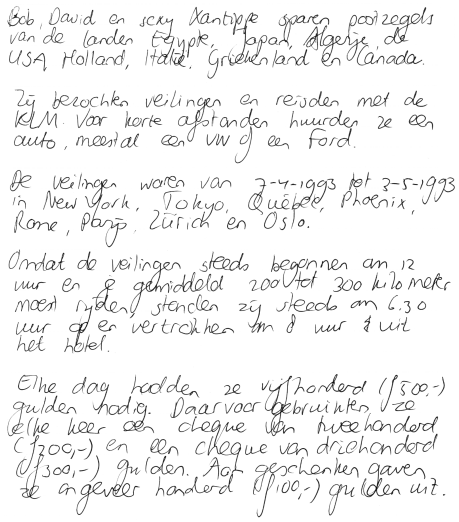
Figura 3.3: Amostras da base IAM [59].

A base *Firemaker* é formada por 251 escritores, a maioria estudantes, na qual cada um escreve, a próprio punho, quatro diferentes documentos (cada estilo em uma página de papel A4), totalizando 1004 páginas. Todas as quatro cartas foram redigidas em Holandês, língua nativa das pessoas que cederam as amostras. Na primeira página, os escritores utilizam um texto base para gerar amostras de texto-dependente (*cópia*). Na segunda página, os escritores observam um desenho animado e descrevem seu conteúdo com suas próprias palavras, gerando texto-independente (*natural*). A terceira página escrita, refere-se ao estilo *caixa alta*, neste caso, há um texto base, assim, todos os escritores possuem a mesma quantidade de texto. Por último, o escritor é obrigado a distorcer sua escrita, a fim desta não ser reconhecida, descrita como *falsificação*. Na *falsificação*, todos os escritores utilizam um texto de apoio, ou seja, é também texto-dependente. Os autores da base descrevem este estilo como falsificação, entretanto perante a ciência forense este estilo é descrito como dissimulação, pois neste caso o escritor dissimula sua escrita. Nesta tese empregaremos o termo falsificação como proposto pelos autores da base para nos referirmos a este estilo. As imagens da base *Firemaker* encontram-se digitalizadas em 256 níveis de cinza com 300 *dpi* em formato *TIFF*.

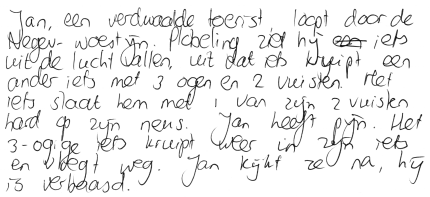
Utilizando a base *Firemaker*, poderemos abordar diferentes técnicas, por exemplo, avaliar o impacto de possuirmos somente texto-dependente ou somente texto-independente, ou ainda avaliar o desempenho quando temos uma mistura (*mix*) dos três diferentes estilos no conjunto de treinamento. Outra característica interessante, desta base, é o fato

de podermos trabalhar com o conceito de falsificação, de modo que, quando o escritor distorce sua escrita, sistemas baseados em características locais poderiam classificar esta distorção como um texto não pertencente ao mesmo escritor.

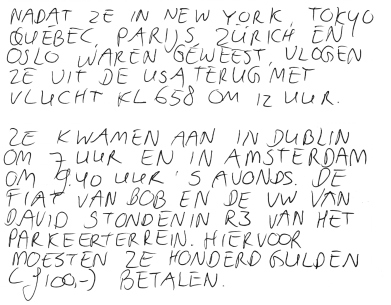
A Figura 3.4 apresenta os quatro exemplos de um mesmo escritor variando o estilo da escrita (texto-dependente ou *cópia*, texto-independente ou escrita *Natural*, *caixa alta* e *falsificação*).



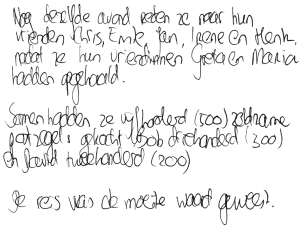
(a)



(b)



(c)



(d)

Figura 3.4: Na Figura (a) temos um exemplo de texto-dependente; na Figura (b) temos uma amostra de texto-independente. Na Figura (c), um exemplo de caixa alta e, por fim, a Figura (d) demonstra uma tentativa de disfarçar a própria escrita, falsificação.

### 3.1.4 Comentários

Atualmente, há um grande número de bases de dados públicas e privadas, empregadas em diversos trabalhos [93]. O número de escritores e amostras colhidos variam bastante [56, 58]. Podemos, ainda, encontrar bases em diversas línguas, tais como: Chinês, Grego, Árabe, Tailandês, Francês, Persa, entre outras [93], cada uma apresentando níveis de dificuldades diferentes. As bases empregadas neste trabalho foram: BFL, IAM e *Firemaker*. Através destas três bases será possível demonstrar a eficiência da abordagem em diferentes línguas escritas e, também no uso de texto-dependente e texto-independente. Outro fator que poderá ser observado é o desempenho em comparação ao número de escritores. A união das três bases pode demonstrar que o uso de textura, como característica, é menos sensível com a variação da língua que outras características alográficas. A Tabela 3.1, a seguir, descreve algumas das bases de dados existentes, o número de escritores, a língua escrita, o tipo (texto-dependente (TD) ou texto-independente (TI)) e a quantidade de amostras existentes.

Tabela 3.1: Comparação entre diferentes bases de dados *off-line*.

| Base | Número de Escritores | Língua | Tipo | Documentos |
|---|---|---|---|---|
| BFL [35] | 315 | Português | TD | 945 |
| IAM [59] | 657 | Inglês | TI | 1539 |
| *Firemaker* [88] | 251 | Holandês | TD e TI | 1004 |
| RIMES [39] | 1300 | Francês | TI | 12723 |
| QUWI [58] | 1017 | Árabe e Inglês | TD e TI | 5085 |
| CEDAR [94] | 1500 | Inglês | TD | 1500 |
| ICFHR-2012 [56] | 100 | Inglês e Grego | TD | 400 |
| Demokritos [56] | 126 | Inglês e Grego | TD | 504 |

## 3.2 Revisão Bibliográfica

Classificaremos os trabalhos, de acordo com a abordagem empregada na extração das características, as quais podem ser classificadas como local e global. É comum o uso de ambas as abordagens, a fim de avaliar o desempenho da junção das mesmas. A abordagem local utiliza detalhes da grafia, fazendo uso de algum método de segmentação. No caso de documentos manuscritos, os grafemas poderiam caracterizar letras ou palavras segmentadas do documento manuscrito. Na abordagem global, partes maiores do documento podem ser utilizadas, como: parágrafos, linhas ou até mesmo o texto como um todo. Alguns autores descrevem estes como atributos genéticos e genéricos [42].

### 3.2.1 Abordagens Locais

Zois e Anastassopoulos [99] descrevem em seu trabalho taxas de acertos na identificação de escritores próximas a 95%, usando palavras. Os autores demonstram que seu sistema é confiável e adaptativo a outras línguas, por meio de duas bases, nos idiomas Inglês e Grego, através de uma base de dados composta por 50 escritores, em que cada escritor escreve 45 amostras da mesma palavra.

Em seus experimentos, as imagens foram fragmentadas entre 5 e 25 partes. Através de uma transformada morfológica de um histograma horizontal da palavra, obteve-se o vetor de características. Dois sistemas de classificação são utilizados: Classificadores Bayesianos e Redes Neurais MLP. Os autores reportam taxas de acertos de 92% usando Redes Bayesianas e 96,5% usando MLP. Em relação à língua escrita (Inglês e Grego), obtiveram-se taxas bastante próximas, comprovando que sua proposta independe da língua escrita.

Baranoski [7] descreve, com precisão, a base BFL e realiza experimentos utilizando características de inclinação axial e distribuição de borda direcional, como métricas, para verificação de escritor. Um vetor de distâncias é gerado, computando a distância euclidiana entre exemplos de manuscritos de referência e exemplos de manuscritos questionados, sendo estes vetores de distâncias, fornecidos como entrada para o classificador SVM. Utilizando 50 escritores para o processo de treinamento e 265 para testes, Baranoski [7] obteve taxas de acertos de 89,6%, (2,5% de erro para falsa rejeição e 7,9% de erros para falsa aceitação).

No trabalho de Bensefia et al. [9] foi proposto um método de verificação e identificação de escritores usando uma abordagem local, apresentando taxas de acertos de 86% para identificação e 96% para verificação. Basicamente, o processo de identificação é tido como um processo de recuperação de informações, empregando o *Vector Space Model* (VSM). Assim, um documento $D$ passa a ser representado por um conjunto de pesos atribuídos às características. Testes variando entre um parágrafo e algumas palavras (três ou quatro) foram realizados. Neste trabalho, foram empregados 150 escritores da base IAM.

Outro trabalho relacionado com a verificação e identificação de escritor é apresentado por Bulacu et al. [20]. No trabalho, os autores utilizam *FDP's* (Função de Distribuição de Probabilidade) para caracterizar a individualidade do escritor. A ideia principal dos autores é realizar uma análise da textura da imagem, observando a forma do caractere, desta maneira, a metodologia utilizada apresenta-se como independente do conteúdo textual da escrita.

Usando a base IAM, em seus experimentos, Bulacu et al. [20], alcançaram taxas de identificação entre 68-89% para Top-1 e 91% à 96% para Top-10. A taxa de acerto em função do Top-$N$ significa que foram computados como acerto, se pelo menos um documento do escritor questionado aparecer em uma lista de tamanho $N$. Em experimentos com verificação, as taxas de acerto ficam entre 91% e 97,2%. Os autores utilizaram 650 escritores em seus experimentos.

Um interessante trabalho utilizando a base *Firemaker* é apresentado por Schomaker et al. [87]. A base *Firemaker* possui um caráter mais desafiador, pois conta com cartas manuscritas com quatro estilos de escrita: Texto-dependente, texto-independente, caixa alta e falsificação, assim, tende a apresentar taxas de acertos mais baixas, quando utilizados todos os estilos. Os autores utilizam um *codebook* de componentes conectados através de contornos fragmentados e apresentam altas taxas de acerto, utilizando texto-dependente, 97%. Entretanto, os melhores resultados alcançados nos experimentos indicaram uma taxa de acerto de 70% para caixa alta, 70% para texto-independente e 50% para falsificações. Os autores mostram que, avaliando o Top-10, houve melhoras nas taxas de acerto de 50% para 90%, no caso mais extremo. Neste trabalho, fica claro que algumas bases e alguns estilos de escrita trazem uma maior dificuldade que outras, sendo poucos os trabalhos que fizeram estudos, neste nível, com a base *Firemaker*. É comum o uso de somente um estilo de escrita, como o texto-dependente ou texto-independente. Nestes experimentos, foram utilizados 150 escritores para testes, empregando mapas de Kohonem auto organizável como método de classificação.

Recentemente, Brink et al. [18] apresentaram um interessante trabalho na identificação de escritor. Para afirmar a eficiência do método proposto, os autores classificaram suas bases em dois grandes grupos: Manuscritos medievais e manuscritos modernos. Para cada grupo foram utilizadas duas diferentes bases, sendo: para a base de manuscritos modernos, as bases IAM e *Firemaker*. Através da abordagem proposta, Brink et al. [18], utilizaram, como informação, a largura dos traços combinadas com a direção, descrevendo como característica *Quill* $p(\emptyset, w)$. Esta característica basicamente captura a relação entre a largura local $w$ e a direção $\emptyset$ dos traços de tinta, considerando uma distribuição de probabilidades. Tal característica consegue definir propriedades da caneta utilizada e ainda resgata características únicas inerentes ao escritor, através das variações de largura dos traçados de tinta. De modo geral, combinando quatro características (contorno do traçado, medidas de ângulo, medidas de largura e cálculo da distribuição de probabilidades) conseguiram um poderoso método para identificação de escritor. As taxas descritas variam de 71% à 97%, considerando Top-1 para bases modernas. Já para Top-10, temos uma variação de 89% à 98%. Brink et al. [18], utilizando somente os estilos *Cópia* e *Natural* da base *Firemaker*, alcançam taxas de 86% de identificação.

Comumente temos visto trabalhos reportando os sucessos alcançados em competições [56], bem como o uso de bases de dados geradas especificamente para competições [66]. Louloudis et al. [56] descrevem os casos de sucesso da competição ICFHR 2012, na qual foram utilizados 100 escritores, empregando texto-dependente nas línguas Grega e Inglês. No trabalho, são descritos os sete melhores métodos, reportando abordagens empregadas, juntamente com as taxas obtidas. Uma série de experimentos foram realizados provando a superioridade de um método sobre o outro. O método Tebessa-c obteve a melhor taxa para Top-1, 94,5%. A menor taxa de acerto foi de 70,3%. Com relação ao Top-10, obtiveram-se taxas de 95,3% à 99,3%, pertencendo também ao método Tebessa-c esta

melhor taxa. Quando avaliado, somente documentos com texto em Inglês, o método Tsinghua obteve resultados superiores, 94,0%. Diversos testes foram realizados a fim de identificar qual método é realmente o mais robusto. O método eleito nesta competição foi a Tebessa-c, já que, em geral, apresentou ligeira vantagem sobre os outros. As características utilizadas no método Tebessa-c baseiam-se em multi-características, as quais se utilizam de pixels pretos do traçado da tinta e do fundo branco para extrair algumas distribuições de probabilidade. No método descrito como mais eficiente, a distância de Manhattan foi utilizada para comparar dois documentos.

Amaral et al. apresentaram, recentemente, três trabalhos empregando a base BFL [2, 4, 3]. Entretanto, utiliza um percentual do número de escritores em cada um deles, 20, 100 e 200 escritores, respectivamente. Utilizando características relativamente simples, como número de linhas escritas contidas na carta, a proporção de pixels pretos, distâncias das margens, entre outras, extrai um vetor de 64 características. Nos três trabalhos, os autores utilizaram o classificador SVM, empregando duas cartas para treinamento e uma para teste. As melhores taxas de acerto, reportadas com 20 escritores no conjunto de teste, são de 80% de acerto na identificação de escritor [2]. Em [4], também reportam taxas de 80%, contudo, nestes experimentos foram empregados 100 escritores no teste, adicionando, como característica, a inclinação axial, aumentando 17 atributos ao vetor de características. Amaral et al. [3], utilizando as mesmas características empregadas anteriormente, avaliam com mais detalhes o esquema de combinação de características, empregando até 200 escritores no teste. Avaliando as características separadamente, percebe-se que estas, isoladamente, apresentam baixas taxas de acerto. Amaral et al. [3] avalia o impacto da quantidade de escritores no conjunto de testes, variando de 20 a 200 escritores. Utilizando a mesma abordagem, demonstra baixas taxas de acerto, 31%, empregando características isoladas. Entretanto, ao combinar características, atinge taxa de 74% de acerto, empregando 200 escritores. Dos oito conjuntos de características empregados, a fusão de três deles (número de linhas da carta, posição da margem inferior e inclinação axial) apresentou o melhor resultado.

Saranya e Vijaya [82, 81] fizeram um estudo minucioso com relação ao classificador SVM e seus parâmetros. Utilizando características relativamente simples, geram um vetor com 26 características em ambos os trabalhos. Utilizando palavras e letras, com o intuito de identificar o escritor, através de uma base contendo 10 escritores e cada documento cedido contendo 100 palavras, Saranya e Vijaya atingem taxas de 94,27% de acerto na identificação. Demonstrando que o *kernel* polinomial apresenta desempenho superior de dois pontos percentuais em relação ao RBF e, aproximadamente, 17 pontos percentuais a mais que o *kernel* linear. Alguns experimentos para estimar valores para os parâmetros $C$, $d$ e $\gamma$ são demonstrados neste trabalho.

### 3.2.2 Abordagens Globais

Através de uma abordagem baseada na análise da textura, Said et al. [79], apresentaram taxas de acertos de 96%. Utilizando uma base contendo 40 escritores, com 25 fragmentos de 128 × 128 extraídos de cada escritor, avaliados em dois conjuntos A e B. No conjunto A, foram utilizadas dez imagens para treinamento e quinze para testes, por escritor; para o conjunto B, quinze imagens para treinamento e dez para testes. Através de multicanais dos filtros de Gabor, foram obtidas dezesseis imagens resultantes, quatro para cada uma das quatro frequências (4, 8, 16 e 32). Ao todo, 32 características foram extraídas, a partir das imagens resultantes. Através da técnica de GLCM (*Gray Level Co-ocurrence Matrix*), cinco distâncias foram avaliadas (d = 1, 2, 3, 4 e 5) em conjunto com as quatro principais direções (0°, 45°, 90° e 135°), usando, para isso, uma imagem binarizada, e extraindo características de Energia, Entropia, Contraste e Correlação para as matrizes GLCM. Para classificação, os autores empregaram classificadores relativamente simples, como, WED (*Weighted Euclidean Distance*) e $k$-NN (*k-nearest neighbor*) , obtendo taxas de 96% de acerto para identificação de escritor. Percebe-se que o classificador WED apresentou melhores resultados, se comparado ao $k$-NN. Outro ponto importante foi que, ao usar somente duas distâncias (d = 1, 2) no GLCM, os autores obtiveram melhores resultados.

Marti et al. [60] usando o classificador $k$-NN e textos de um pequeno número de escritores da base IAM, apresentaram taxas de acerto de 90,7%. Para isto, utilizaram um esquema de zoneamento, dividindo cada linha da amostra em três zonas (zona superior, inferior e conteúdo do meio), extraindo doze características estruturais para cada zona.

Shen et al. [25], utilizaram técnicas 2-D *Gabor Wavelet* para extração de características globais. Através de um método semelhante ao usado por Hanusiak et al. [42], os autores fazem uma normalização, eliminando espaços em branco existentes entre linhas e palavras. Segundo os autores, isto passa a caracterizar a imagem como uma textura mais proeminente. Em seus experimentos, foram realizadas análises com algumas técnicas de extração de textura, como: *Multichannel Decomposition* (MCD), *Line-Based Spectrum Resolution* (LBSR) e GLCM. Os filtros de Gabor Multicanal apresentaram melhor desempenho que as outras características. No processo de classificação, foi usado o classificador $k$-NN. Com uma base proprietária, formada por 50 escritores, com fragmentos possuindo tamanho de 128 × 128, uma taxa média de identificação de 97,6% foi reportada.

He e Tang [44] utilizaram manuscritos Chineses para identificação de escritor, analisando a imagem de forma textural e empregaram técnicas de filtro Gabor para o processo de extração de características. Entretanto, a ideia principal de seu trabalho foi avaliar, por meio de experimentos, o desempenho das abordagens dependente de escritor e independente de escritor, na combinação dos métodos. Usando um classificador WED e uma base formada por 100 manuscritos Chineses, cedidos por 50 escritores, taxas de acertos de 42% (Top-1) e 97% (Top-10) foram descritas.

Imdad et al. [47], usando um pequeno número de escritores da base IAM, 30, reportaram taxas de acerto próximas a 83%. Para tal, os autores utilizaram DHT (*Discrete Hermite Transform*) como características e o SVM como classificador. Em seus experimentos, utilizaram cinco linhas de cada amostra do escritor. Contudo, devido ao número de escritores utilizados ser reduzido, fica difícil avaliar seu real desempenho.

Schlapbach et al. apresentam dois interessantes trabalhos com HMM (*Hidden Markov Model*) [84] e GMM (*Generalized Markov Model*) [85], empregando características geométricas na verificação e identificação de escritor. Para isto, 100 escritores da base IAM foram utilizados, juntamente com nove características geométricas para cada linha da amostra. Em seus experimentos, Schlapback et al. [84], reportaram taxas de 96% para identificação e 97,5% para verificação. Através do uso de GMM, Schlapbach et al. [85] demonstraram um ganho de quase 2% em relação ao seu trabalho anterior, na identificação. A taxa de acerto apresentada por Schlapbach et al. em [85], apresenta-se como uma das melhores taxas de identificação de escritor descrita em literatura, entretanto, os autores utilizaram uma fração da base IAM, ficando difícil saber se estas taxas se mantêm ao utilizar os 650 escritores da base.

Outro trabalho usando textura para identificação de escritor é descrito por Kurban et al. [55]. Neste trabalho, os autores propõem o uso de Gabor e ICA (*Independent Component Analysis*). A imagem da textura é, primeiramente filtrada por conjunto de filtros de Gabor e, em seguida, vetores de características das dimensões mais elevadas foram construídos a partir das texturas filtradas. Uma seleção e redução de dimensionalidade das características foi realizada através de PCA (*Principal Component Analysis*) e algoritmos genéticos. Desta forma, através de ICA, os vetores já com dimensão reduzidas foram extraídos. Usando o classificador $k$-NN em seus experimentos com um $k = 5$, conseguiram taxas que vão de 83,4% nos experimentos usando Gabor e 96 características, até 92,5% com 20 características, usando Gabor, PCA e ICA.

Garain et al. [36], apresentaram um método para representação de escritores através de um conjunto $2D$ Auto-Regressivo (AR). Utilizando as bases RIMES e ISI (*Handwritten Character Databases of Indic Scripts*) obtiveram resultados interessantes para Top-10, 96,5%. Entretanto, para Top-1, as taxas descritas não foram expressivas: 62,1%. Para o processo de identificação de escritor foram computados os coeficientes AR para cada amostra dos escritores da base. Sendo $\widehat{\theta_i}$ a estimativa de coeficientes AR para o $i$-ésimo escritor. Assim, para uma amostra desconhecida, o coeficiente AR é calculado $(\widehat{\theta})$. O próximo passo consistiu em calcular a distância Euclidiana entre o exemplo em questão e algum exemplo contido na base, através de: $d(\widehat{\theta}, \widehat{\theta_i}) = \left\|\widehat{\theta} - \widehat{\theta_i}\right\|^2$. A $j$-ésima amostra será considerada pertencente ao escritor, caso a condição $d(\widehat{\theta}, \widehat{\theta_j}) > d(\widehat{\theta}, \widehat{\theta_j}) \forall i, i \neq j$ seja verdadeira.

Herrera-Luna et al. [45] descrevem alguns experimentos realizados com 30 escritores para identificação de escritor em texto manuscrito. Cada escritor cede três exemplos do

mesmo texto, sendo que o conteúdo variou de cinco a nove linhas. Na abordagem proposta por Herrera-Luna et al. [45], foi realizada uma modificação no algoritmo de classificação supervisionada ALVOT (Algoritmos de Voto). As características usadas referem-se a linhas ou palavras. Em relação à linha, destacam-se: espaços existentes nas margens direita e esquerda, espaço entrelinhas, e espaços entre palavras. Já em relação à palavra têm-se: proporção central da palavra, zonas altas, zonas baixas, inclinação e crista. No total, 22 características foram extraídas, demonstrando taxas de acerto que variaram entre 88,89% para Top-1 e 94,44% para Top-$q$ (o valor de $q$ não é descrito).

Um trabalho avaliando o desempenho do descritor de textura GLCM, em diferentes domínios de aplicações, é apresentado por Martins et al. [61]. Neste trabalho, os autores focam somente na verificação de escritor. Utilizando abordagem proposta por Hanusiak et al. [42], atingem taxas de 94,49% de acerto na verificação. Para tal, foram utilizados 115 escritores no conjunto de teste e, 200 no treinamento. Usando um único descritor, dos quatorze propostos por Haralick et al. [43], o descritor de Energia, variando a distância de 1 até 5, combinados com os ângulos 0°, 45°, 90° e 135°, geram um vetor de 20 características. Experimentos apresentados por Martins et al. são interessantes, pois descrevem a robustez do descritor de textura GLCM para diferentes texturas.

Bertolini et al. [12], empregando a abordagem proposta por Hanusiak et al. [42], demonstra que novos descritores de textura podem ser ainda mais robustos no processo de verificação e identificação de escritor. Empregando conceitos de dissimilaridade proposto por Pekalska e Duin [72], demonstra que a abordagem é promissora, relatando taxas de acerto de 99,2% empregando a base BFL (texto-dependente) e 96,7% para base IAM (texto-independente), no processo de identificação de escritor. Foram utilizados dois descritores de textura LBP (*Local Binary Pattern*) e LPQ (*Local Phase Quantization*), demonstrando a superioridade dos mesmos, quando comparados ao GLCM, empregado no trabalho de Hanusiak [42] e Martins [61]. O descritor LPQ apresentou vantagem sobre o LBP, quando o número de referências utilizadas eram pequenas ($R = S = 3$). Utilizando 9 referências, percebe-se que a diferença nas taxas de identificação entre os dois descritores diminuem. Outra lacuna deixada no trabalho de Hanusiak foi o impacto do tamanho dos blocos de textura. Os escritores observaram este fato e notaram que blocos muito pequenos ($64 \times 64$) podem reter pouca informação do escritor. Em geral, pôde-se observar que blocos maiores apresentam melhores resultados, entretanto, em bases com texto-independente, não há como assegurar a geração de blocos grandes com vasto conteúdo. Contudo, através dos experimentos, fica claro que blocos com dimensões $256 \times 128$ ou $256 \times 256$ apresentam desempenho igual aos blocos de $768 \times 256$. Experimentos com diferentes regras de fusão são descritos no trabalho, demonstrando que a regra da soma, na maioria dos casos, apresentam os melhores resultados. O impacto do número de escritores no conjunto de treinamento também foi avaliado. Através destes experimentos pôde-se observar que, aumentando o número de escritores no conjunto de treinamento, é possível atingir resultados satisfatórios, mesmo com poucas referências. Por fim, foi feita

uma comparação entre diferentes estratégias de classificação (*Pairwise*, um contra todos e dissimilaridade), demonstrando o ótimo desempenho da abordagem de dissimilaridade.

Djeddi et al. [29] realizaram um interessante estudo sobre reconhecimento de escritor, utilizando texto-independente. Para isso, utilizaram bases de dados disponibilizadas para competições, contendo manuscritos em Latim e Grego. A base de dados é composta por 126 escritores, os quais contribuem com quatro páginas, duas em Latim e duas em Grego. Observando a grafia nas duas diferentes línguas, percebe-se a grande diferença entre as mesmas. As características utilizadas baseiam-se no conjunto de informações extraídas de Matrizes GLRL (*Gray Level Run Length*). Um vetor de 600 características é extraído através do GLRL modificado, detalhes de execuções e ângulos utilizados são descritos no trabalho. Djeddi et al. comparam o método proposto com trabalhos de outros escritores, demonstrando as características empregadas e a dimensão do vetor de cada um deles. Para base em Língua Grega, seu método apresentou as melhores taxas 92,06%, utilizando o classificador SVM. Dentre os dez métodos comparados, utilizando a língua Inglesa, o método proposto também alcançou a melhor taxa, 87,30%. Empregando ambas as línguas, os escritores reportam taxas de 76,59% de acerto na identificação de escritor. No processo de verificação, os autores reportam taxas médias de erros EER (*Equal Error Rate*) de 2,78% (Grego), 3,57% (Inglês) e 5,75% (Ambas). A abordagem proposta por Djeddi et al. [29], em todos os casos, apresentaram resultados superiores aos demais. Um fato interessante é a língua Grega ter apresentado resultados melhores que a língua Inglesa, segundo os autores, isso se deve ao fato dos escritores envolvidos possuírem, como língua nativa, o Grego. Outro fato é que a abordagem proposta por Siddiqi e Vicent [92] apresentaram taxas bem inferiores para esta base de dados, talvez isto se deva aos experimentos descritos pelos mesmos serem realizados unicamente com texto-dependente.

Um trabalho recente apresentado por Newell e Griffin empregando a base IAM com fins de avaliar o descritor de Características Básicas da Imagem Orientadas a Colunas ($oBIF$) é apresentado em [65]. Neste trabalho, os autores fazem um comparativo entre o descritor oBIF e outros citados em competições internacionais como ICDAR e ICFHR. Empregando somente escritores que possuem duas "passagens", como descrevem os autores, resulta em um conjunto de 301 escritores da base IAM. Newell e Griffin reportam taxa de 99% de acerto na identificação de escritor utilizando todo o texto e 89,6% utilizando uma única linha do texto. A abordagem oBIF descrita pelos autores foi utilizada, anteriormente, em reconhecimento de caracteres e em reconhecimento de texturas, sendo adaptada, aqui, para a identificação de escritor. Todos os detalhes da abordagem oBIF são descritos com riqueza de detalhes em [65]. Uma questão levantada pelos autores nesse trabalho é sobre competições. Comumente temos visto excelentes taxas de acertos em competições que possuem vasto conhecimento e capacidade neste tipo de organização. Entretanto, é de conhecimento de todos que equipes participantes maximizem o desempenho do seu próprio método para determinado conjunto de dados. Tal atitude não é ilegal, entretanto, para tal

conjunto de dados, o método pode ser ótimo, com mérito de prêmio; para outro conjunto, pode ser que tal método não seja satisfatório. Um tópico interessante, que os autores discutem, refere-se a métodos alográficos e métodos baseados em textura. Os argumentos dos autores são quanto à comparação letra a letra em alguns métodos alográficos ficarem restritos a uma determinada língua. Os autores discutem, ainda, diversos tópicos, abrindo discussões, demonstrando também que o descritor oBIF, sendo empregado em uma base de competição de língua Árabe, apresenta boas taxas de acerto, 93,1% com 204 escritores. Em todos os experimentos foi utilizado um classificador de vizinho mais próximo com a distância euclidiana.

### 3.2.3 Combinação de Abordagem Local e Global

Srihari et al. [94], cuja motivação se deve a algumas sentenças em tribunais americanos, iniciam seu trabalho de maneira a demonstrar a hipótese de que o manuscrito é individualista. Usando técnicas similares às usadas por peritos, os autores avaliaram o uso de algumas características básicas para determinar quantitativamente a individualidade em técnicas de aprendizagem de máquina. Srihari et al. [94], usaram macrocaracterísticas com objetivo de capturar características globais, a partir de um documento manuscrito, e microcaracterísticas com intuito de analisar características locais, como a forma do caractere. Taxas para identificação de escritor de 81% foram descritas, sendo que a hipótese de validar o problema da individualidade foi atingida.

Schalapbach e Bunke [83], apresentaram bons resultados para identificação e verificação de escritor 96,56% e 97,5%, respectivamente, utilizando a base IAM e um classificador baseado em HMM (*Hidden Markov Model*). A partir de uma linha arbitrária da amostra de texto, dada como entrada, cada classificador HMM obtém uma pontuação de reconhecimento. Desta maneira, foram extraídas nove características de cada grade: número de pixels pretos, centro de gravidade, momento de segunda ordem, posição e direção do contorno da parte superior e inferior, número de transições preto para branco, e a fração de pixels entre a parte superior e inferior.

Em sua tese, Pecharromán [5] realizou um estudo com verificação e identificação de escritor, utilizando textura. Para tal, o autor descreve ter usado cinco diferentes características, extraídas a partir de: contorno, imagem pré-processada sem ruídos, imagem binária e imagens em tons de cinza. Destas características, seis possuem foco em FDP's e a característica de autocorrelação horizontal. As taxas para verificação ficaram próximas a 96,5%, quando combinadas as três características. Já para identificação, Pecharromán [5] obteve taxas de 82,04%, considerando o Top-1 e, 93,61% para Top-10. As taxas, combinando as mesmas três características, também apresentaram melhores resultados para o processo de identificação.

Em seu recente trabalho, Siddiqi et al. [92], fizeram uma interessante pesquisa em verificação e identificação de escritor, utilizando duas bases conhecidas na área: IAM

e RIMES. Em seus experimentos, foram utilizadas 15 características, sendo classificadas como características globais, locais e baseadas em polígonos. Características locais apresentaram piores resultados quando isoladas. As baseadas em polígono apresentaram melhores taxas. Combinando as características, os autores obtiveram melhor resultado que quando usadas isoladamente. Um interessante experimento foi o de avaliar diferentes quantidades de amostras, como: uma palavra, duas palavras, uma linha até quatro linhas, e a página completa. Entretanto, a ideia principal deste trabalho foi verificar a presença de padrões redundantes na escrita dos escritores e também o uso de atributos visuais. Através de experimentos realizados sobre as duas bases, Siddiqi et al. [92] notaram a eficácia de sua proposta. Os autores perceberam que usando uma ou duas palavras, as taxas são inferiores a 50%. As taxas permaneceram próximas ao variar a quantidade de conteúdo utilizado, quatro linhas até todo o conteúdo. Após um procedimento de seleção de características, os autores observaram que, das quinze características utilizadas, seis são indispensáveis, duas são parcialmente relevantes e sete são irrelevantes. Quanto aos resultados obtidos, Siddiqi et al. [92] apresentaram excelentes resultados na identificação para a base IAM, 91% e 97% (Top-1 / Top-10) e 97,77% para verificação, através da combinação de características.

Önder e M. Bilginer [51] em seu trabalho com identificação de escritor, apresentaram excelentes resultados em relação a trabalhos similares. Usando a base IAM com 93 escritores, os autores reportam taxa de acerto de 98,76%. Em seus experimentos, os autores utilizaram características globais e locais e três classificadores foram avaliados: $k$-NN, GMM e NDDF Bayes (*Normal Density Discriminant Function*), sendo o último, responsável pelas melhores taxas. As características globais, isoladamente, apresentaram resultados pouco interessantes. A combinação das características locais apresentou melhores resultados que quando combinadas as características globais e locais. Um estudo interessante, realizado pelos autores, a respeito da base IAM, demonstrou que menos da metade dos escritores apresenta mais de 12 linhas escritas em suas amostras.

Embasado no trabalho de Baranoski [7], Hanusiak et al. [42], propõem a geração de um conteúdo textural a partir de uma técnica de eliminação de espaços em branco entre linhas e palavras. Usando seis, dos quatorze descritores propostos por Haralick et al. [43] e inclinação axial, os autores realizaram um interessante trabalho de verificação de escritor, utilizando a base BFL. A fim de verificar a influência do número de escritores no desempenho do modelo (por se tratar de um modelo escritor-independente), diferentes conjuntos de treinamento foram avaliados, cada um variando o número de escritores em (25, 50, 100 e 200). Em seus experimentos, com 115 escritores para o conjunto de testes, são descritas taxas de acertos de 96,1%.

Brink et al. [17] apresentam um estudo sobre a correção da inclinação não natural no processo de verificação de escritor. Através de dois experimentos, buscaram verificar a real importância da característica inclinação. Os autores descrevem que a característica inclinação, em sistemas biométricos, é muito menos relevante do que se supunha, pois,

com a eliminação da inclinação natural, percebe-se uma queda de 1 a 5 pontos percentuais nas taxas. Demonstram também que, mesmo corrigindo a inclinação, foi possível alcançar um melhor desempenho em relação a sua eliminação, entretanto, apresentou taxas bem abaixo que usando a escrita normal. Este trabalho é interessante, já que tantos trabalhos são realizados buscando a correção da inclinação ou a eliminação da mesma. Muitos outros buscam extrair a inclinação da palavra. Bertolini et al. [11] usando assinaturas, mostra que a característica inclinação apresentou-se como uma ótima característica quando combinada com outras.

## 3.3 Considerações Finais

Em suma, a avaliação de trabalhos publicados, ao longo dos anos, contribui fortemente na elaboração desta tese, pois é possível perceber a evolução dos sistemas de verificação e identificação de escritor. Todavia, como descrito anteriormente, realizar um estudo comparativo entre resultados obtidos, considerando as abordagens utilizadas para o processo de verificação e identificação, torna-se difícil devido à diversidade de bases de dados existentes e também ao emprego de diferentes números de escritores. Na Tabela 3.2, apresentaremos um resumo dos trabalhos descritos anteriormente. Assim, podemos observar, de forma simples, o desempenho destes, através dos anos e também as características empregadas: bases, número de escritores e o método de classificação.

Tabela 3.2: Síntese da revisão bibliográfica

| Ref. | Base | Ano | Característica | Escritores | Classificação | Desempenho (%) | |
|---|---|---|---|---|---|---|---|
| | | | | | | Verificação | Identificação |
| Baranoski et al. [7] | BFL | 2005 | Inclinação Axial | 315 | SVM | 89,6 | - |
| Hanusiak et al. [41] | BFL | 2010 | GLCM - Inclinação Axial | 115 | SVM | 96,1 | - |
| Martins et al. [61] | BFL | 2011 | GLCM | 115 | SVM | 94,4 | - |
| Amaral et al. [2] | BFL | 2012 | 64 Características | 20 | SVM | - | 80,0 |
| Amaral et al. [4] | BFL | 2013 | 64 Características | 100 | SVM | - | 80,0 |
| Amaral et al. [3] | BFL | 2013 | 64 Características + Inclinação | 200 | SVM | - | 74,0 |
| Bertolini et al. [12] | BFL | 2013 | LBP e LPQ | 115 | SVM | 99,4 | 99,2 |
| Marti e Bunke [60] | IAM | 2001 | Características Estruturais | 20 | k-NN | - | 90,7 |
| Schlapbach et al. [83] | IAM | 2004 | Características Geométricas | 120 | HMM | 97,5 | 96,5 |
| Bensefia et al. [9] | IAM | 2005 | *Codebook* de Características | 150 | VSM | 96,0 | 86,0 |
| Bulacu et al. [20] | IAM | 2007 | *Codebook* de Características | 650 | Dist. Hamming | 97,2 | 89,0 |
| Imdad et al. [47] | IAM | 2007 | Direcionais de Hermite | 30 | SVM | - | 83,0 |
| Schlapbach et al. [84] | IAM | 2007 | Características Geométricas | 100 | HMM | 97,5 | 96,0 |
| Balbás et al. [5] | IAM | 2007 | PDF's + Auto-correlação | 650 | - | 96,5 | 82,0 |
| Schlapback et al. [85] | IAM | 2008 | Características Geométricas | 100 | GMM | - | 97,8 |
| Siddiqi et al. [92] | IAM | 2010 | *Codebook* de Características | 650 | Dist. X2 | 97,7 | 91,0 |
| Kirli et al. [51] | IAM | 2011 | Características Globais e Locais | 93 | NDDF | - | 98,7 |
| Brink et al. [18] | IAM | 2011 | Característica *Quill* | 657 | - | - | 97,0 |
| Bertolini et al. [12] | IAM | 2013 | LBP e LPQ | 240 | SVM | 99,6 | 96,7 |
| Newell et al. [65] | IAM | 2013 | oBIF | 301 | Distância Euclidiana | - | 99,0 |
| Schomaker et al. [87] | *Firemaker-Cópia* | 2007 | *Codebook* de Características | 150 | *Kohonen* | - | 97,0 |
| Schomaker et al. [87] | *Firemaker-Natural* | 2007 | *Codebook* de Características | 150 | *Kohonen* | - | 70,0 |
| Schomaker et al. [87] | *Firemaker-Caixa Alta* | 2007 | *Codebook* de Características | 150 | *Kohonen* | - | 50,0 |
| Schomaker et al. [87] | *Firemaker-Falsificação* | 2007 | *Codebook* de Características | 150 | *Kohonen* | - | 50,0 |
| Brink et al. [18] | *Firemaker* | 2011 | Característica *Quill-Hinge* | 251 | Vizinho Mais Próximo | - | 86,0 |
| Said et al. [79] | - | 1998 | Gabor e GLCM | 40 | WED | - | 96,0 |
| Zois et al. [99] | - | 1999 | Características Morfológicas | 50 | MLP | - | 96,5 |
| Srihari et al. [94] | - | 2002 | Micro e Macro Características | 1000 | k-NN | - | 81,0 |
| Cong et al. [25] | - | 2002 | Características de Textura | 50 | k-NN | - | 97,6 |
| Zyhe et al. [44] | - | 2004 | Gabor | 50 | WED | - | 97,0 |
| Kurban et al. [55] | - | 2009 | Gabor e ICA | 55 | k-NN | - | 92,5 |
| Garain et al. [36] | RIMES+ISI | 2009 | Conjunto AR $2D$ | 422 | - | - | 62,1 |
| Herrera-Luna et al. [45] | - | 2011 | Características de Linha e Palavras | 30 | ALVOT | - | 88,8 |
| Louloudis et al. [56] | ICFHR 2012 | 2012 | Multi-características | 100 | Dist. Manhattan | - | 94,5 |
| Saranya et al. [82, 81] | - | 2013 | Características Locais | 10 | SVM | - | 94,2 |
| Djeddi et al. [29] | Demokritos | 2013 | Matriz GLRL | 126 | SVM | 94,2 | 76,5 |

# CAPÍTULO 4

# MÉTODO PROPOSTO

Neste capítulo apresentaremos o método proposto para o desenvolvimento desta tese. A Figura 4.1 apresenta uma ideia geral dos procedimentos realizados. Descreveremos, a seguir, cada uma das etapas em detalhes.

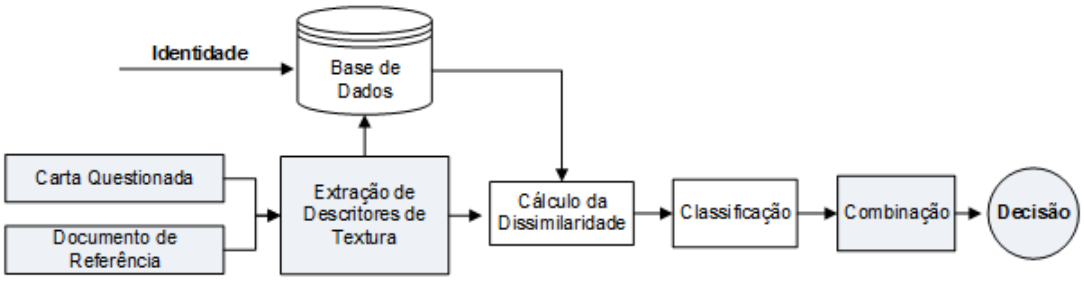



Figura 4.1: Modelo proposto para identificação e verificação de escritor.

Conforme a abordagem proposta, apresentada na Figura 4.1, através de diferentes bases de dados, aplicou-se o modelo proposto por Hanusiak et al. [42] com intuito de gerar, por meio da escrita, um rico conteúdo textural. Em seguida, diferentes descritores de textura foram avaliados, tais como: GLCM, LBP e LPQ. Empregamos a abordagem escritor-independente, juntamente com a dissimilaridade, visando, principalmente, à independência dos modelos. Assim, em caso de inserção de novos escritores no conjunto de teste, não há necessidade de treinar novamente o modelo. Inicialmente, nossos experimentos utilizavam conjuntos de treinamento com números fixos de escritores. O propósito era observar a robustez do modelo, de acordo com o número de escritores. Em todos os experimentos empregamos o classificador SVM (*Support Vector Machine*) devido ao mesmo ser descrito como robusto, em trabalhos com duas classes. Desta maneira, a partir das probabilidades geradas pelo classificador SVM, em relação ao número de vetores de dissimilaridades na representação do escritor, podemos avaliar diferentes regras de decisão para combinar estas saídas. A decisão final será gerada a partir da conclusão das etapas anteriores [12]. Verificamos o desempenho desta abordagem na identificação de diferentes estilos de escrita, como: texto-dependente, texto-independente, caixa alta e a falsificação. Para falsificação, na qual o escritor distorce sua escrita a fim de se passar por outra pessoa, a abordagem apresentou-se bastante promissora [13]. Por fim, propomos um método inovador, o qual seleciona escritores para compor o conjunto de treinamento, na tentativa de gerar modelos robustos através de um número limitado de escritores. Neste caso, empregaremos Algoritmos Genéticos, a fim de selecionar escritores para fazer parte do conjunto de treinamento. Percebemos, então, que todas as etapas têm grande importância para que esta aplicação obtenha sucesso.

## 4.1 Bases de Dados

Com o propósito de compararmos as taxas alcançadas com as apresentadas por Hanusiak et al. [42], decidimos manter o mesmo protocolo em relação ao número de escritores nos conjuntos de treinamento e teste. Desta forma, dos 315 escritores da base BFL, 115 foram utilizados no conjunto de testes. Quatro diferentes partições para treinamento foram consideradas com os escritores restantes, cada uma com 25, 50, 100 e 200 escritores. A ideia era verificar o impacto do número de escritores usados para treinamento do desempenho global. Desta forma, avaliamos o quão impactante é o número de escritores no conjunto de treinamento para o desempenho do sistema.

Na literatura, encontramos grandes diferenças no número de escritores empregado nos trabalhos, utilizando a base IAM. Como exemplo, Marti et al. [60] utiliza 30 escritores, enquanto Brink et al. [18] emprega toda a base. De acordo com a abordagem escritor-independente, escritores que fazem parte do conjunto de treinamento não estão presentes no conjunto de testes, assim, não foi possível utilizar toda a base. Desta forma, resolvemos utilizar os mesmos percentuais do protocolo aplicado à base BFL. Dos 650 escritores existentes nesta base, 240 foram utilizados para testes. Os escritores restantes também foram divididos em quatro partições, com: 50, 100, 205 e 410 escritores.

A fim de compararmos os resultados publicados por Schomaker et al. [87], utilizamos a base *Firemaker* nas seguintes proporções, dos 250 escritores existentes na base, 150 foram usados no conjunto de teste. Os 100 escritores restantes, foram subdivididos em quatro conjuntos de treinamento, cada qual contendo 20, 40, 80 e 100 escritores. Entretanto, para os experimentos com seleção de escritores, necessitamos de dois conjuntos validação, assim, especificamente nestes experimentos, utilizamos 90 escritores no conjunto de teste e continuamos com as mesmas quantidades nos subconjuntos de treinamento. Lembramos que a base *Firemaker* é formada por quatro diferentes estilos, sendo que temos por escritor: uma carta com texto-dependente, uma carta com texto-independente, uma carta com escrita em caixa alta e uma carta na qual o escritor é obrigado a disfarçar sua escrita, gerando assim, uma falsificação. Desta forma, avaliaremos cada estilo de grafia separadamente.

Uma união das três bases (BFL, IAM e *Firemaker*) foi realizada a fim de obtermos uma base diversificada com um vasto número de escritores. A partir desta união, temos documentos escritos em três diferentes línguas: Português, Inglês e Holandês. Com isso, podemos analisar o impacto da geração de um modelo criado com escritores de diferentes línguas, pois, ao utilizar textura para representar a escrita de um escritor, é possível obter características indiferentes da língua escrita, diferente do uso de características alográficas. A Tabela 4.1 apresenta uma síntese da quantidade de escritores usados para teste e treino em cada base.

Tais partições, com diferentes números de escritores, deram início aos nossos experimentos; assim, a proposta deste trabalho é selecionar escritores para compor o conjunto

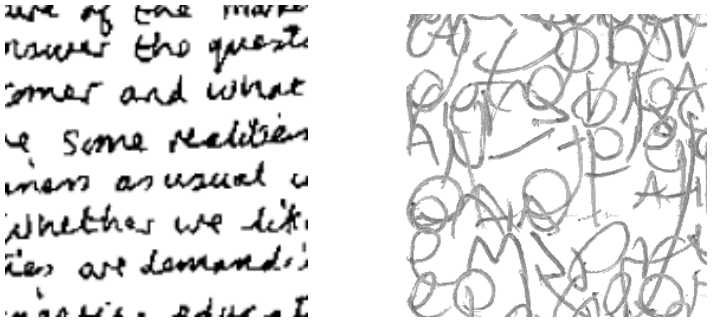
Tabela 4.1: Quantidade de escritores nos conjuntos de treinamento e teste.

| Base | Teste | Treino 1 | Treino 2 | Treino 3 | Treino 4 |
|---|---|---|---|---|---|
| BFL | 115 | 25 | 50 | 100 | 200 |
| IAM | 240 | 50 | 100 | 205 | 410 |
| *Firemaker* | 150/90 | 20 | 40 | 80 | 160 |
| BFL + IAM + *Firemaker* | 440 | 95 | 190 | 385 | 775 |

de treinamento, e não, utilizar uma partição com um número fixo de escritores, como demonstrado anteriormente. Desta forma, este estudo inicial do desempenho, empregando diferentes números de escritores será necessário para analisarmos o desempenho ao possuirmos mais e menos escritores no conjunto de treinamento. Logo após, foi possível avaliar a abordagem proposta de seleção de escritores.

## 4.2 Geração do Conteúdo Textural

Said et al. [79] demonstram que avaliar a escrita como textura, sem nenhum pré-processamento não é um método eficiente. Isto, devido ao fato da textura ser afetada pelos diferentes espaços existentes entre linhas e palavras, gerando uma grande quantidade de espaços em branco e pouco conteúdo de textura, propriamente dita. Assim, Said et al. [79] utilizaram um algoritmo de normalização com o intuito de minimizar estes fatores. Embasado nisto, Hanusiak et al. [42] propõem um método para eliminação de espaços em branco entre palavras e linhas, mantendo as características do escritor. Desta maneira, o texto redigido seria incompreensível, mas a textura formada seria forte para representação de um escritor. As Figuras 4.2(a) e 4.2(b) representam imagens utilizando as abordagens propostas por Said et al. [79] e Hanusiak et al. [42], respectivamente.



(a) (b)

Figura 4.2: Na Figura 4.2(a) temos um exemplo gerado a partir da abordagem usada por Said et al. [79]. Já a Figura 4.2(b), representa a abordagem proposta por Hanusiak [42].

Neste trabalho, utilizamos a abordagem proposta por Hanusiak et al. [42], já que a

mesma demonstrou-se mais robusta. Nessa abordagem, todo conteúdo original é transposto, sendo eliminada grande parte dos espaços em branco. A presença de espaços em branco é de certa forma, uma característica do escritor; entretanto, a grande quantidade de espaços em branco existentes nas amostras, pode não ser interessante ao empregarmos descritores de textura. Esta abordagem é totalmente independente da língua empregada.

Descreveremos, a seguir, o método de geração de textura, o qual se inicia com um processo de segmentação, de forma a separar segmentos conexos (presença de momentos ou espaçamentos), preservando uma característica importante: seu ângulo de inclinação original. Assim, o processo reorganiza os segmentos encontrados em um novo alinhamento, elimina espaços em branco entre as linhas e entre as palavras, e ainda entre segmentos de palavras. Para este processo, sumarizamos as etapas do algoritmo proposto para realizar o método de eliminação de espaços em branco em documentos manuscritos.

1. **Pré-processamento:** A amostra original é binarizada e armazenada;
2. **Busca:** A imagem é percorrida de cima para baixo e da esquerda para a direita, até encontrar o primeiro componente de traço (pixel preto);
3. **Marcação:** São marcados na imagem todos os pixels conexos ao pixel encontrado. Utiliza-se o preenchimento de área (*fill area*) [42] como lógica de preenchimento, aplicado de forma recursiva e observando as oito direções em relação ao pixel encontrado;
4. ***Buffer:*** O conteúdo a ser extraído é marcado com outra cor na imagem binarizada (em memória);
5. **Transferência:** A partir de dados de altura e largura, busca na imagem original o conteúdo conexo, transcrevendo-o para uma imagem auxiliar (Figura 4.3);

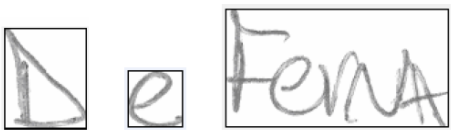


Figura 4.3: Componentes selecionados pelo algoritmo de preenchimento de área.

6. **Comparação:** Realizamos uma comparação entre o conteúdo binarizado com o conteúdo original, a fim de selecionar apenas o texto correspondente ao grupo preenchido pelo *fill area*. O objetivo é eliminar traços recortados junto ao *bounding box*, mas que não fazem parte do conteúdo selecionado pelo preenchimento de área (Figura 4.4);
7. **Nova Imagem:** Realiza a transferência da imagem original para uma nova imagem, de forma a rearranjar todos os pixels obtidos pela imagem binária. Como linha de base, utilizamos o ponto médio do conjunto extraído; (Figura 4.5);

Figura 4.4: Exemplo de recorte do *bounding box* contendo componentes não selecionados pelo preenchimento de área.

Figura 4.5: Componentes selecionados dispostos lado a lado.

8. **Finaliza-Inicializa:** É extraído um conjunto de pixels da imagem binarizada, iniciando a busca por uma nova ocorrência, repetindo os passos.

Desta maneira, serão gerados os conteúdos texturais, a partir das bases de dados. Na Figura 4.6, podemos observar um exemplo de textura gerado a partir da carta de um escritor da base BFL.

Figura 4.6: Amostra do conteúdo de textura gerado a partir de uma carta manuscrita.

Uma característica observada por Hanusiak et al. [42] é a sobreposição existente. A distância entre as linhas de base são determinadas pela média das alturas dos componentes da linha corrente. Contudo, escritores que apresentam em sua grafia uma maior ocorrência de ascendentes e descendentes com proporcionalidade baixa são representados por blocos menores. O objetivo dessa análise é entender melhor qual é a importância da densidade da textura.

Na Figura 4.7, apresentamos um exemplo da carta original da base IAM e, ao lado, os fragmentos de textura utilizados em nossos experimentos.

### 4.2.1 Densidade da Textura Gerada

O método de geração de textura proposto por Hanusiak et al. [42] apresenta vantagens perante o uso de imagens originais. Entretanto, não sabemos se o método como a textura é gerada poderia originar texturas que expressassem melhor as características do escritor, caso a densidade fosse outra. Por exemplo, uma sobreposição entre os componentes conexos reduziria ainda mais o conteúdo de espaços em branco, podendo proporcionar uma textura ainda mais densa. Ou também, com a sobreposição de componentes, podemos perder informações relevantes do escritor, de maneira que, aumentando a densidade, possamos estar empobrecendo a representação do escritor através da textura.

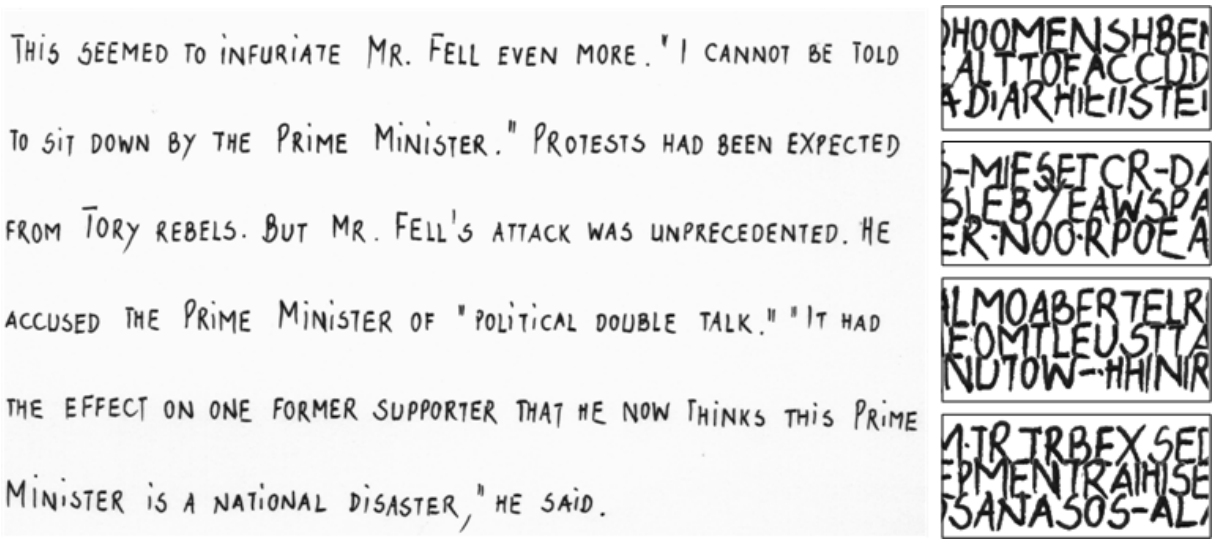
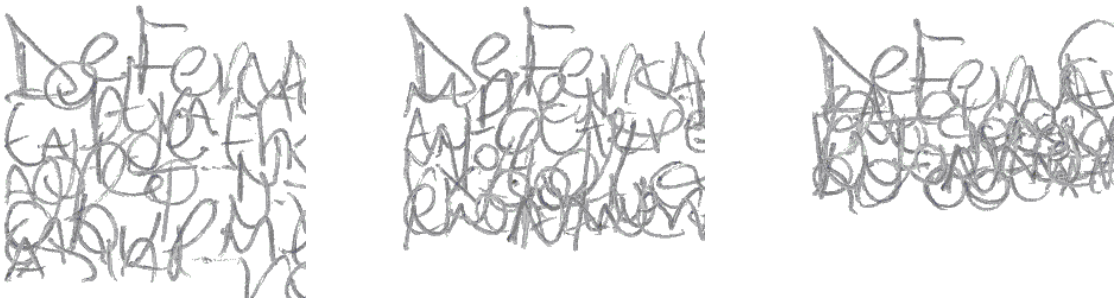



Figura 4.7: Carta original e blocos de textura.

Assim, propomos avaliar componentes conexos mais e menos sobrepostos, a fim de observar se existe um impacto nas taxas de verificação e identificação de escritores. A Figura 4.8 mostra blocos de textura utilizando a abordagem e os parâmetros empregados por Hanusiak et al. [42] e com componentes mais sobrepostos gerando uma textura mais densa. A Figura 4.8(a) apresenta a abordagem utilizada por Hanusiak et al. [42], utilizando uma média para gerar os fragmentos de textura. Embasados nesta abordagem, reduzimos em 10 e 25 % as distâncias das larguras e alturas entre os componentes, de maneira que houvesse uma sobreposição de 10 e 25% maior que a abordagem original. A Figura 4.8(b) apresenta uma redução de 10% na altura e 10% na largura. Já a Figura 4.8(c) mostra uma diminuição de 25%, tanto para a altura quanto para a largura.



(a) (b) (c)

Figura 4.8: Sobreposição de componentes conexos: Figura (a) padrão original utilizado por Hanusiak et al. [42], Figura (b) compactação 10% maior que a original e Figura (c) compactação de 25% maior que a original.

### 4.2.2 Dimensão dos Fragmentos de Textura

Said et al. [79] e Hanusiak et al. [42] utilizam em seus trabalhos amostras de tamanhos diferentes, 128 × 128 e 256 × 256, respectivamente. Entretanto, nenhum estudo foi realizado para demonstrar o impacto do tamanho do fragmento no desempenho do sistema. Desta forma, avaliamos o impacto da dimensão do bloco de textura na verificação de escritores.

Algumas características têm impacto direto na dimensão e no número de fragmentos, por exemplo, a quantidade de texto escrito. Bases como BFL e *Firemaker Cópia*, na qual temos amostras de texto-dependente, podemos assegurar que todos os escritores possuem a mesma quantidade de texto escrito. As alterações referem-se ao tamanho da letra e proporções do espaço que a escrita ocupa, porém, tais detalhes são comportamentos que servem como características dos mesmos. Para as bases IAM e *Firemaker Natural* com texto-independente, não podemos garantir se a quantidade de texto escrito pelo escritor é suficiente para gerar conteúdo textural que seja significativo. Isto porque o número de linhas escritas varia bastante, principalmente na base IAM, sendo assim, é impossível avaliar todas as dimensões para ambas as bases.

Como avaliamos o impacto da quantidade de fragmentos na geração de vetores de dissimilaridade (*vide* Seção 4.4), variando de 1 a 9, é necessário que, através do documento escrito, seja possível gerar no mínimo nove blocos de textura. Desta forma, propomos avaliar três diferentes alturas para base BFL, (64, 128 e 256 pixels), e sete diferentes larguras (64, 128, 209, 256, 329, 460 e 768).

Devido à escrita, às vezes, serem muito pequenas ou ainda escritores que escrevem de maneira compacta, podemos ter blocos de textura com um vasto conteúdo de espaços em branco, pois não é possível preencher, com textura, toda a dimensão do bloco. Um fator a ser estudado é o impacto ao possuirmos espaços em branco na parte superior e inferior do bloco (pois o conteúdo é centralizado). Tal comportamento é mais comum ao utilizar 256 pixels de altura. A Figura 4.9 ilustra dois tamanhos de blocos de textura bastante empregados na literatura, 128 × 128 e 256 × 256 pixels.

## 4.3 Descritores de Textura

A literatura apresenta uma grande quantidade de trabalhos utilizando abordagens estatísticas e espectrais [42, 79, 25, 55, 44] e um menor número utilizando características estruturais [97, 30]. Desta forma, avaliaremos descritores como: GLCM, LBP e LPQ para as abordagens estatística e estrutural. Assim, poderemos observar a robustez dos descritores para esta aplicação.

A partir do trabalho de Hanusiak et al. [42] resolvemos explorar detalhes do GLCM, os quais não foram levados em consideração. Devido aos sistemas de identificação serem mais críticos que sistemas de verificação, desejamos verificar se as taxas apresentadas por

(a) (b)

Figura 4.9: Na Figura (a) temos um exemplo de bloco de textura de dimensão 128 × 128 pixels. A Figura (b) apresenta um bloco de 256 × 256 pixels.

um descritor na verificação de escritor mantém-se para identificação, e ainda analisar o quão robusto é cada método para o processo de identificação. Avaliaremos também se, na abordagem de seleção de escritores, o descritor tem forte influência.

### 4.3.1 GLCM

Said et al. [79] descrevem experimentos utilizando GLCM, variando o parâmetro de distância. Os autores demonstram em seu trabalho que distâncias mais baixas (1 e 2) apresentaram resultados melhores, contudo, Hanusiak et al. [42], em seus experimentos, adotam o uso de cinco distâncias (1, 2, 3, 4 e 5), contrariando os fatos apresentados por Said et al. [79].

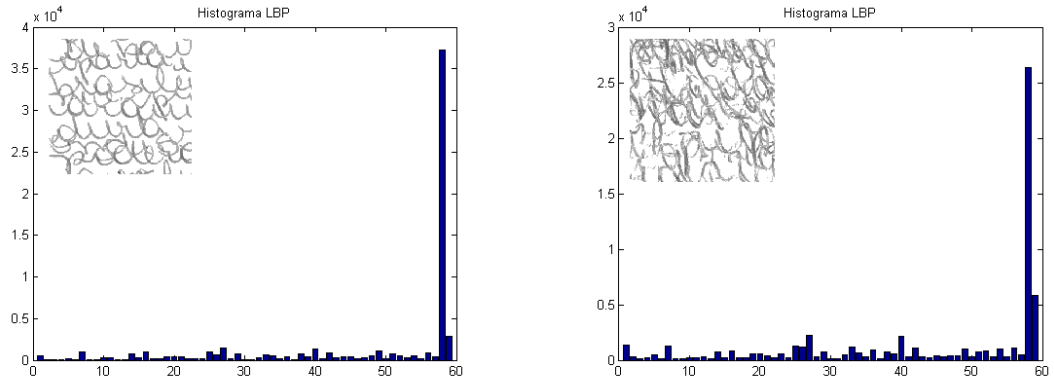
Desta maneira, percebemos a necessidade de avaliar o impacto do uso de diferentes distâncias no descritor GLCM. Outra proposta seria combinar algumas das medidas descritas por Haralick [43]. Das quatorze medidas propostas por Haralick, quatro são comumente citadas: Energia, Contraste, Correlação e Homogeneidade. Assim, a ideia é avaliar estas medidas isoladamente e combinadas. O vetor de características extraído através do descritor GLCM varia de tamanho quatro, quando utilizamos somente uma medida, aliada à distância um, (considerando as quatro principais direções) até tamanho 16 quando, ao invés de utilizar uma única medida, concatenamos as quatro medidas citadas anteriormente.

O método proposto na década de 70, ainda é bastante empregado e, em geral, apresenta bons resultados. Sendo este descritor bastante utilizado como referência em aplicações com textura, poderemos utilizá-lo como base de comparação para novos descritores.

### 4.3.2 LBP

Uma proposta para abordagem estrutural é o *Local Binary Pattern* [64], que vem apresentando resultados promissores em diversas áreas, [97, 53, 30]. O LBP, assim como o GLCM, possui alguns parâmetros básicos a serem avaliados. Os dois principais são o $P$, que representa o número de vizinhos e o $R$, que se refere ao Raio. Atualmente, existem variações do descritor LBP que são: LBP uniforme $LBP^{U2}_{P,R}$, LBP invariante à rotação $LBP^{RI}_{P,R}$ e o LBP uniforme, invariante à rotação $LBP^{RIU2}_{P,R}$. Nossa proposta é verificar se os parâmetros descritos como proeminentes na literatura se mantêm nesta aplicação.

Conforme descrito na seção 2.1.2.1, o vetor gerado pelo LBP uniforme $LBP^{U2}_{P,R}$ possui 59 características. Desta forma, propomos realizar análises quanto ao histograma LBP, já que nesta aplicação verifica-se a existência de uma grande quantidade de pixels brancos nas imagens. A Figura 4.10 apresenta duas imagens utilizadas neste trabalho, juntamente com seu histograma LBP. Percebe-se, então, que a ocorrência da transição branco-branco é bastante representativa, mesmo após o processo geração de textura. Assim, temos o objetivo de observar o impacto ao reduzir a dimensão do histograma LBP, removendo contagens de transições como a posição 58 do histograma.




(a) (b)

Figura 4.10: Histograma LBP.

### 4.3.3 LPQ

Ojansivu e Heikkilä [69] demonstram que, mesmo o foco do descritor LPQ ser para imagens borradas, a técnica apresenta ótimos resultados para imagens sem ruídos deste tipo. Um dos parâmetros básicos a serem avaliados é o tamanho da janela para esta aplicação. Nosso foco para esta abordagem é analisar a robustez deste descritor, direcionado para imagens borradas, aplicado ao problema de identificação de escritor. Utilizaremos, neste trabalho, o histograma gerado pelo descritor LPQ como vetor de características para classificação, possuindo 256 valores. Experimentos preliminares avaliando diferentes ta-

manhos de janelas como, ($3 \times 3$, $5 \times 5$, $7 \times 7$, $9 \times 9$, $11 \times 11$), demonstraram que janelas de tamanho $7 \times 7$ apresentam melhor desempenho.

## 4.4 Dissimilaridade

Neste trabalho, utilizaremos o conceito de dissimilaridade (*vide* seção 2.2). Na abordagem de dissimilaridade, o número de vetores de dissimilaridade gerado é totalmente dependente do número de blocos de textura extraídos por escritor, ou seja, o número de referências. A Figura 4.11 apresenta o conceito da abordagem de dissimilaridade usando amostras de um mesmo escritor e amostra de diferentes escritores.

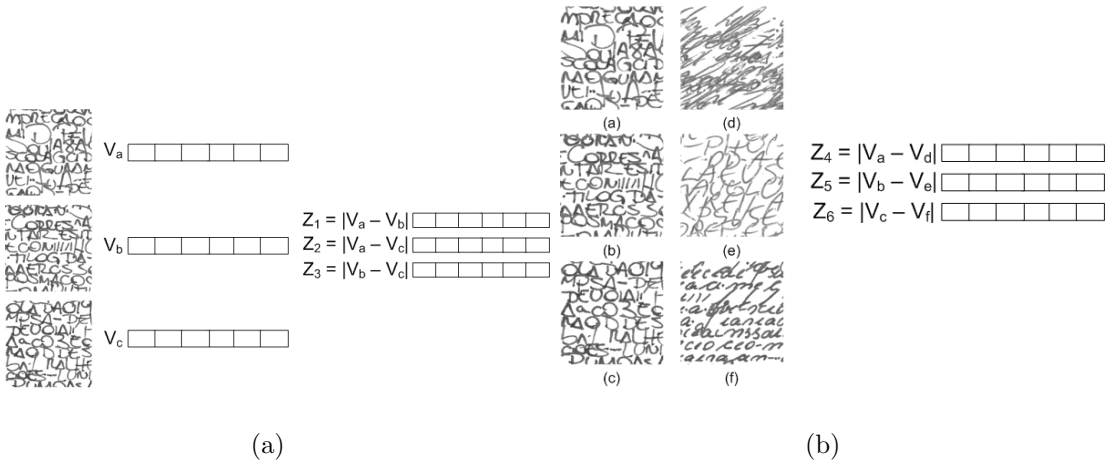



Figura 4.11: Vetores de dissimilaridade gerados a partir dos vetores de características:(a) Dissimilaridade entre amostras do mesmo escritor; (b) dissimilaridade entre amostras de escritores diferentes.

Podemos notar na Figura 4.11 que, para três amostras de um escritor, através da combinação entre as três, podemos gerar três vetores de dissimilaridade positivos. Para manter o sistema balanceado, geramos a mesma quantidade de vetores de dissimilaridade negativos (*vide* Algoritmo 1). Contudo, o número de blocos de textura disponíveis é bastante limitado devido ao conteúdo escrito por escritor ser inconstante para bases de texto-independente. A dimensão do bloco também implica diretamente na quantidade de blocos de textura a serem gerados. Assim, resolvemos avaliar o impacto utilizando 3, 5, 7 e 9 amostras (blocos de textura) por escritor, já que é possível obter até nove amostras para todos os escritores das três diferentes bases. A Tabela 4.2 apresenta, para as diferentes quantidades de amostras utilizadas, o número de vetores de dissimilaridades positivo e negativo computados.

O estudo do número de amostras (referências) necessárias para criarmos bons modelos têm dois motivos principais. O primeiro deve-se ao fato de validarmos a proposta demonstrando que esta é robusta, tendo pouco ou muito conteúdo escrito. O segundo motivo é que podemos avaliar diferentes regras de combinações, avaliando as probabilidades de

Tabela 4.2: Vetores de dissimilaridade gerados em relação à quantidade de amostras por escritor.

| Número de Amostras | Vetores Positivos | Vetores Negativos | Vetores por Escritor |
|---|---|---|---|
| 3 | 3 | 3 | 6 |
| 5 | 10 | 10 | 20 |
| 7 | 21 | 21 | 42 |
| 9 | 36 | 36 | 72 |

saída geradas pelo classificador SVM. Como podemos variar o número de referências, tanto no conjunto de treinamento como no conjunto de testes ($R$ e $S$), pretendemos explorar o impacto destes em ambos os conjuntos.

Em conjunto com o número de referências, um objeto de estudo deste trabalho é o número de escritores no conjunto de treinamento. Diversos trabalhos têm intuito de demonstrar que, aumentando a quantidade de escritores no conjunto de treinamento, é possível melhorar as taxas de verificação ou identificação de escritores [12, 42, 13].

Investigamos aqui o quão impactante é a quantidade de escritores para gerarmos modelos robustos. De acordo com os trabalhos [12, 13], pudemos perceber que o número de escritores no conjunto de treinamento é menos importante do que se supunha. Assim, para cada base de manuscritos a ser empregada, utilizaremos uma proporção para o conjunto de testes e o restante será utilizado no conjunto de treinamento. A fim de observar tais detalhes, geramos quatro diferentes partições, utilizando os $W$ escritores não empregados no conjunto de testes, assim temos: $(i)$ todos os $W$ escritores; $(ii)$ metade dos $W$ escritores ($\frac{1}{2}$); $(iii)$ um quarto dos $W$ escritores ($\frac{1}{4}$) e $(iv)$ um oitavo dos $W$ escritores ($\frac{1}{8}$).

Motivados com isso e percebendo que esta análise é necessária para a abordagem futura de seleção de escritores, desenvolvemos parte deste trabalho fundamentado na proposta de utilizar um número fixo de escritores para gerarmos modelos.

## 4.5 Diferentes Estilos de Escrita

Inicialmente, investigamos o desempenho do processo de identificação de escritores utilizando texto-dependente e texto-independente. Percebemos que, devido à quantidade uniforme de texto escrito em bases de texto-dependente, foram alcançados melhores desempenhos, comparado ao uso de texto-independente. Possivelmente esta diferença no desempenho deve-se aos blocos de textura que contêm mais informações do escritor.

Também, mais dois outros estilos foram investigados neste trabalho: o estilo caixa alta e a falsificação. Estes estilos são exclusivos da base *Firemaker* (*vide* Seção 3.1.3). Em nossos experimentos, avaliamos duas propostas. Na primeira, utilizamos o mesmo estilo no conjunto de treinamento e no conjunto de teste. Na segunda abordagem empregamos

uma mistura de estilos (texto-dependente, texto-independente e caixa alta) avaliando se os modelos gerados poderiam ser ainda mais robustos.

Desta forma, em nossos experimentos, foi avaliado o desempenho na identificação de escritores utilizando três estilos de escrita (texto-dependente, texto-independente e caixa alta), além da falsificação. Como no processo de falsificação o escritor é instruído a se passar por outra pessoa, em sistemas de verificação e identificação não é empregado o estilo falsificação no processo de treinamento, já que em casos reais não é possível conseguir amostras de falsificações com antecedência.

## 4.6 Seleção de Escritores

Um dos propósitos deste trabalho foi avaliar o impacto do número de escritores presente no conjunto de treinamento para geração de um modelo. Para isso, inicializamos nossos experimentos subdividindo o número de escritores restantes, aqueles não usados no conjunto de testes (*vide* Seção 4.4). Desta forma, foi possível observar o desempenho do modelo gerado através dos subconjuntos de treinamento, contendo diferentes números de escritores. Nesta abordagem, observamos se a conjectura de quanto mais escritores no conjunto de treinamento, melhor, é verdadeira ou falsa para aplicações escritor-independente.

A princípio, nosso objetivo foi avaliar se, além dos escritores presentes no conjunto de treinamento, os blocos de textura de cada escritor poderiam influenciar no desempenho do sistema. Como os blocos de textura são escolhidos aleatoriamente, e alguns poderiam ser mais indicados para o processo de dissimilaridade, viu-se a hipótese de investigar tal propriedade. Entretanto, observando os blocos de textura gerados através de manuscritos de um mesmo escritor, notamos uma uniformidade de textura entre eles. Ao selecionar amostras aleatórias de um determinado escritor para gerar os vetores de dissimilaridade, verificamos que o desvio padrão nas taxas de identificação era baixo. A partir disso, nossa abordagem focou unicamente na seleção de escritores. A Figura 4.12 demonstra este comportamento uniforme entre amostras do mesmo escritor. Em aplicações que exista uma alta variação intraclasse, pode ser interessante a seleção de amostras, juntamente com a seleção de classes (neste trabalho, a classe representa o escritor) para compor o conjunto de treinamento. A partir da eliminação da necessidade de selecionar amostras, concentraremos nossos esforços no processo de seleção de escritores.

A Figura 4.13 apresenta a distribuição dos conjuntos de dados empregados na abordagem de seleção de escritores, juntamente com detalhes do uso de Algoritmos Genéticos. Podemos perceber que dois conjuntos de validação e um de teste são utilizados. Empregamos três diferentes subconjuntos para provar a robustez do modelo gerado, demonstrando que o modelo é generalista e não especialista para um determinado conjunto. Após a geração dos três subconjuntos de escritores *Validação 1*, *Validação 2* e *Teste* criados aleatoriamente, geramos um subconjunto de escritores para compor o conjunto de treinamento (formado pelos escritores remanescentes). Para os diferentes experimentos descritos nesta

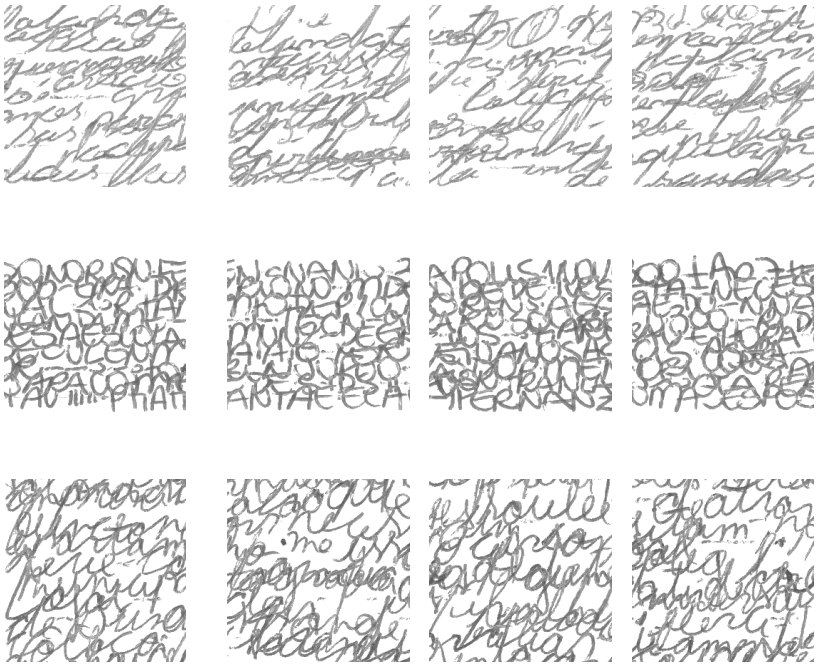


Figura 4.12: Exemplos de uniformidade de textura intraclasse.

tese, os escritores nos conjuntos de *Teste* serão os mesmos, assim, será possível comparar o desempenho de diferentes abordagens. O conjunto de *Validação 1* foi utilizado para avaliar o valor de aptidão do algoritmo de busca. Através deste subconjunto, foi possível avaliar o desempenho do sistema a partir de um conjunto de escritores selecionados do conjunto de treinamento. O subconjunto de *Validação 2* é uma prova de conceito, demonstrando que otimizando o subconjunto *Validação 1* podemos conseguir melhorias em outro subconjunto. Ou seja, o conjunto de escritores selecionados no conjunto de treinamento não é específico para o subconjunto de *Validação 1*. Por fim, será avaliado o modelo gerado em um subconjunto de *Teste*, demonstrando ou não, que selecionar escritores para gerar um modelo pode aumentar o desempenho do sistema.

Os escritores presentes no conjunto de treinamento têm impacto direto nos vetores de dissimilaridade gerados. Amostras do mesmo escritor são utilizadas para representar o escritor, representando os vetores de dissimilaridade positivos. Ao selecionarmos escritores para compor o conjunto de treinamento, implicitamente estamos selecionando amostras, pois, selecionando um escritor para que entre suas amostras sejam gerados vetores de dissimilaridade, a aplicação irá selecionar escritores que gerem um vetor de dissimilaridade positivo e negativo que colaborem para a formação de um modelo robusto. Para gerar vetores de dissimilaridade negativo, necessitamos de amostras de diferentes escritores. Desta forma, a abordagem de seleção de escritores guiada por uma função de aptidão, seleciona escritores a partir de um conjunto de treinamento, buscando maximizar esta função de aptidão. Neste trabalho, empregamos a Taxa de Acerto Global e a Área Abaixo da Curva (*Area Under the Curve - AUC*) como medida de aptidão no Algoritmo Genético.

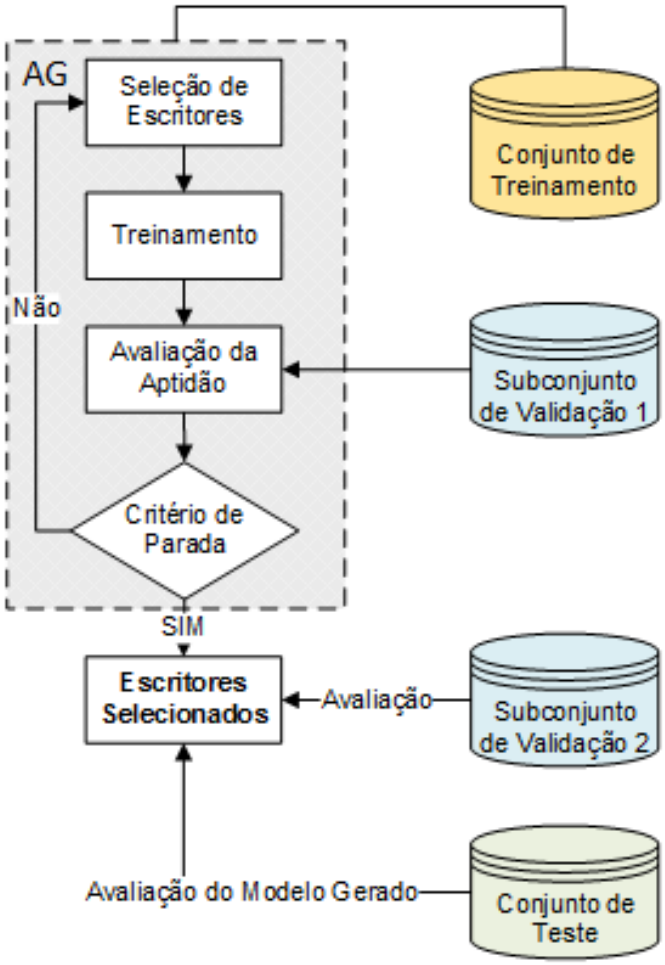



Figura 4.13: Abordagem escritor-independente proposta para seleção de escritores.

Nos conceitos de representação de dissimilaridade, temos o intuito de maximizar as distâncias entre vetores de dissimilaridade positivos e negativos. Buscamos encontrar vetores positivos que sejam os mais próximos de zero. Desta forma, amostras do escritor que tenham uma maior similaridade poderiam contribuir para o desempenho. Entretanto, não sabemos, até o momento, se tais vetores de dissimilaridade contribuem para a criação de modelos robustos, ou ainda, se tivermos amostras do mesmo escritor com um grau de diferença mais elevado, estas amostras seriam melhores para geração deste modelo ideal? Isto se repete para a escolha dos escritores e na geração de vetores de dissimilaridade negativos, uma vez que desejamos vetores de dissimilaridade negativos o mais distante de zero. Entretanto, na criação de modelos não se sabe qual o impacto de utilizar vetores com altas distâncias.

Através dos escritores selecionados, podemos observar se existe uma correlação entre eles, ou seja, analisaremos se existem características visuais que justifiquem a seleção destes escritores. Inicializamos nossos experimentos com uma base sintética com o objetivo de validar a proposta de seleção de escritores. Logo em seguida, empregaremos as seguintes bases de manuscritos: BFL, IAM e *Firemaker*.

A Figura 4.14 mostra um conjunto de treinamento com seis escritores, cada escritor

com três amostras. Através da abordagem proposta, ao invés de empregar todos os seis escritores para gerar um modelo, se selecionássemos (através de um algoritmo de busca) somente três para compor o novo conjunto de treinamento (escritores selecionados em cinza), o modelo gerado por estes seria melhor, pior ou similar? Através de experimentos, poderemos avaliar se há melhoras no desempenho do sistema e o impacto da redução do número de escritores no conjunto de treinamento.

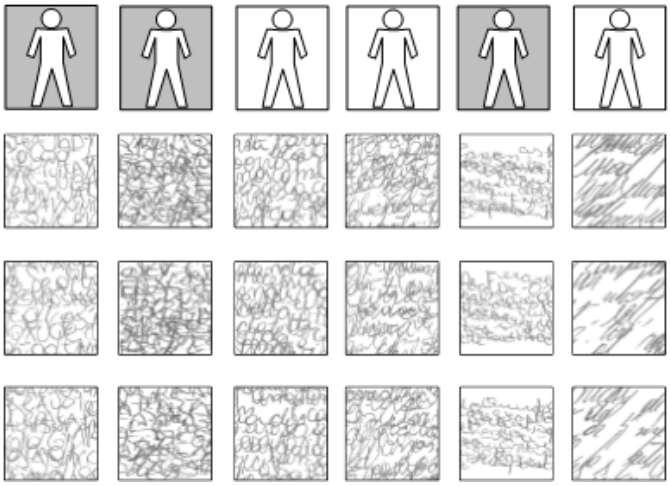


Figura 4.14: Esquema de seleção de escritores utilizando seis escritores com três amostras cada.

### 4.6.1 Algoritmo de Busca

Na abordagem proposta, empregaremos Algoritmos Genéticos com intuito de selecionar escritores para compor o conjunto de treinamento. AG's apresentam-se como um método eficiente em diversas áreas, nas quais o foco é a busca, seleção ou otimização. O uso de AG's pode ser justificado devido a apresentarem bons resultados em problemas similares, cujo foco é selecionar características [98, 86]. Desta forma, optamos por empregar Algoritmos Genéticos com a finalidade de selecionar escritores para gerar um modelo robusto. Nos experimentos iniciais, estimamos alguns parâmetros do AG, como: número de gerações, operações genéticas, condições de parada que seriam realizadas.

Avaliaremos duas diferentes funções de aptidão na busca de minimizar a taxa de erro global ou maximizar alguma medida de diversidade. No contexto de identificação e verificação de escritores, duas métricas são bastante utilizadas para representar o desempenho do sistema, a Taxa de Acerto Global e Área abaixo da Curva (AUC).

A primeira função de aptidão a ser empregada é o cálculo da minimização da taxa de erro global, sendo esta uma métrica bastante utilizada para medir o desempenho do sistema. Ao utilizar esta métrica como função de aptidão, nosso objetivo é melhorar as taxas de identificação com base nos escritores selecionados. A segunda medida a ser empregada é a AUC, representada por um valor escalar que descreve o desempenho esperado. Segundo Fawcett [34], a AUC possui uma importante propriedade estatística,

sendo que a área abaixo da curva de um classificador é a probabilidade que o classificador expressa um exemplo positivo mais altamente aleatório escolhido, do que um exemplo negativo escolhido aleatoriamente.

# CAPÍTULO 5

# EXPERIMENTOS

Nossos experimentos encontram-se divididos em três partes. A primeira parte (*vide* seção 5.1) relata experimentos quanto à verificação e identificação de escritores utilizando abordagem de escritor-independente, cujo principal objetivo é avaliar o impacto dos descritores de textura, o impacto da quantidade de escritores e o número de referências nos modelos gerados. As bases de dados BFL e IAM foram utilizadas nesta primeira etapa, sendo uma base de texto-dependente e outra de texto-independente. Na segunda parte (*vide* seção 5.2) empregamos a base *Firemaker* com o intuito de avaliar os descritores de textura quanto à falsificação e diferentes estilos de escrita, como: texto-dependente, texto-independente, escrita caixa-alta e falsificações. Por fim, a terceira parte (*vide* seção 5.3), descreve a abordagem de seleção de escritores. Nesta seção, apresentamos experimentos através da abordagem de seleção de escritores, empregando uma base de dados sintética e utilizando três diferentes bases de manuscritos, BFL, IAM e *Firemaker*.

Adotamos o classificador SVM (*Support Vector Machine*) em todos os nossos experimentos. No processo de treinamento para validação cruzada foram usados 5-$fold$. Avaliamos diferentes *kernels*, entretanto, melhores resultados foram alcançados utilizando o *kernel* Gaussiano. Parâmetros $C$ e $\gamma$ foram determinados através de busca gulosa. A taxa de acerto é descrita por:

$$Taxa\ de\ Acerto\ Global = 100 - \frac{FP + FN}{TP + TN + FP + FN} \times 100 \tag{5.1}$$

em que, FP, FN, TP e TN representam Falso Positivo, Falso Negativo, Verdadeiro Positivo e Verdadeiro Negativo, respectivamente. A Figura 5.1 descreve a matriz de confusão, a partir das quatro possíveis situações.

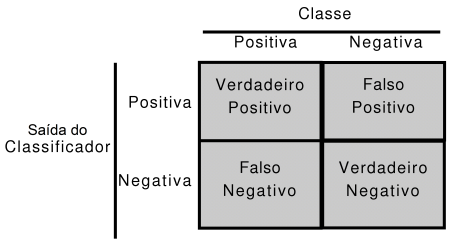



Figura 5.1: Matriz de confusão 2 × 2 representando as quatro situações possíveis.

Além da Taxa de Acerto Global, utilizamos curvas ROC - (*Receiver Operating Characteristic*) [34], juntamente com a Área Abaixo da Curva ROC - (AUC) já que são métricas bastante difundidas para análise de desempenho em problemas com duas classes. Neste trabalho, as estimativas de probabilidades dadas pelo classificador SVM são essenciais

para avaliarmos diferentes estratégias de fusões, como: Voto Majoritário, Máximo, Soma e Mediana. O SVM tem se apresentado como um classificador robusto para produzir probabilidades *a posteriori*, $P(classe|entrada)$. O uso de estimativas de probabilidades, combinadas com regras de decisão, são comumente utilizadas [10, 42]. Neste trabalho, adotaremos a estratégia proposta por Platt [74].

Para realização dos experimentos, utilizamos um conjunto de treinamento e um conjunto de teste. Como aplicamos o conceito de escritor-independente neste trabalho, escritores presentes no conjunto de treinamento não fazem parte do conjunto de testes. Desta forma, foram selecionados, de forma aleatória, escritores para compor cada conjunto. Com o intuito de analisar a robustez do método proposto, utilizamos um esquema de validação cruzada com três conjuntos de testes e três conjuntos de treinamento, assim, nossos resultados são representados pelas taxas médias de acerto.

Devido ao uso de Algoritmos Genéticos na abordagem de seleção de escritores, cada experimento será realizado três vezes, garantindo assim, uma diversidade quanto aos escritores selecionados. Desta maneira, podemos expressar os resultados através das médias das taxas de acertos globais, juntamente com o desvio padrão, permitindo avaliar melhor a dispersão entre os resultados de cada execução.

## 5.1 Avaliação dos Descritores e Parâmetros do Sistema Escritor-independente

Na abordagem escritor-independente, temos o intuito de avaliar o impacto ao possuirmos diversos conjuntos de treinamento, cada qual com diferentes números de escritores. A abordagem, além de ser comumente empregada [42, 12, 62] e apresentar bons resultados, irá nos auxiliar em análises futuras, pois, poderemos observar se, utilizando um grande número de escritores no conjunto de treinamento, é possível construir modelos mais robustos, ou então, se, com um número reduzido de escritores temos taxas de acertos consideráveis. Diferentes descritores de textura também foram avaliados, como GLCM, LBP e LPQ. Outra característica relevante avaliada foi o número de amostras (referências) utilizadas na abordagem de dissimilaridade para os conjuntos de treinamento ($R$) e teste ($S$). Experimentos utilizando diferentes regras de fusão, como Voto Majoritário, Soma, Máximo, Mediana, entre outros, foram avaliados. Apresentamos as Taxas de Acerto Global, alcançadas de acordo com a abordagem, ou seja, quanto à verificação ou identificação de escritores.

### 5.1.1 Verificação de Escritores

Conforme apresentado no Capítulo 3, percebemos que a grande maioria dos trabalhos focam seus esforços no processo de identificação de escritores, não na verificação. Entretanto, muitas das vezes necessitamos somente verificar se determinado documento foi

realmente escrito por certa pessoa, não havendo necessidade de identificação do escritor. Devido a isso, nosso objetivo é demonstrar que nossa proposta é eficiente, tanto para o processo de verificação, quanto para a identificação.

Taxas de acerto referentes ao processo de verificação são sempre melhores que as taxas para identificação, isso por se tratar de uma comparação $1:1$. Experimentos realizados, neste trabalho, apresentaram excelente desempenho para processo de verificação de escritores tanto para base BFL (texto-dependente), quanto para a base IAM (texto-independente).

Verificação de escritor é a tarefa de determinar se um texto manuscrito foi ou não escrito por uma determinada pessoa. Como podemos perceber, trata-se de um problema de natureza binária, de forma que, a partir de um vetor de características $\mathbf{x}$, extraído de um documento $t$ e, uma identidade $I$, determinar se $(I, \mathbf{x})$ faz parte da classe $\omega_1$ ou $\omega_2$. Desta forma, a classe $\omega_1$ indica que a identidade é verdadeira, ou seja, o documento $t$ foi escrito pelo escritor $I$. Por outro lado, a classe $\omega_2$, indica que a identidade é falsa, ou seja, o documento é de um impostor. A fim de comparar os resultados obtidos por Hanusiak et al. [42], utilizamos o mesmo protocolo empregado pelo autor, entretanto, exploramos alguns detalhes não abordados no trabalho deste.

Nos experimentos iniciais realizados com a base BFL, utilizamos 100 escritores para treinamento e 115 para testes. Para geração de vetores de dissimilaridade, cinco referências no conjunto de teste ($S = 5$) e cinco para o conjunto de treinamento ($R = 5$) foram utilizadas. A regra de decisão para combinar as probabilidades *a posteriori* é a regra da soma, pois esta apresenta boas taxas em diversas aplicações [10, 42]. Os fragmentos utilizados na base BFL possuem tamanhos de $256 \times 256$ pixels.

Um interessante estudo realizado por Hanusiak et al. [42], foi o impacto da quantidade de tons de cinza, utilizado na extração de descritores de Haralick et al. [43]. Percebe-se que, para a base utilizada, com apenas dois tons de cinza podemos obter ótimas taxas de acertos. Desta forma, nossos experimentos com GLCM utilizam imagens em dois tons de cinza. Em seus experimentos, os autores utilizam distâncias 1, 2, 3, 4 e 5 concatenadas, isto para diferentes descritores de GLCM e utilizando quatro direções básicas, $0°$, $45°$, $90°$ e $135°$.

Nossos experimentos com GLCM têm o intuito de verificar duas características não exploradas por Hanusiak et al. [42]. A primeira é avaliar o impacto das distâncias no descritor GLCM; e a segunda, é avaliar o uso de diferentes números de referências no processo de treinamento, ou seja, $R = 3, 5, 7$ e 9, fixando o número de referências no teste em cinco ($S = 5$).

Hanusiak et al. [42] apresentam taxas de 95% de acerto através dos parâmetros citados anteriormente. Em nossos experimentos, verificamos que o uso das distâncias, como proposto por Hanusiak et al., apresentam melhores resultados que usando distâncias 1, 2, 3, como proposto por Said et al. [79]. Um segundo experimento foi realizado avaliando quatro diferentes descritores de Haralick et al. [43]. A literatura descreve quatorze

diferentes descritores de Haralick, no entanto, grande parte é correlacionada. Decidimos avaliar quatro destes descritores para verificar o impacto destes, isoladamente (Energia, Contraste, Correlação e Homogeneidade) e quando combinados [95]. Ao combinar as quatro medidas, obteve-se taxa de acerto de 98,26%, taxa superior de quando utilizadas isoladamente. Podemos perceber que a complementaridade entre os descritores é benéfica, gerando uma maior robustez para o sistema. Três regras de combinação das saídas dos classificadores foram utilizadas, sendo elas: Voto majoritário, soma e máximo. Tais regras foram empregadas já que segundo Kittler et al. [52] apresentam ótimo desempenho mesmo existindo ruídos nos dados. Em todos os casos, a regra da soma apresentou melhor desempenho.

Realizamos um experimento com o objetivo de avaliar o desempenho do sistema em relação ao número de amostras no processo de dissimilaridade para o conjunto de treinamento ($R$). Assim, avaliamos se, aumentando o número de referência no conjunto de treinamento ($R$), era possível melhorar o desempenho do sistema. Neste experimento, fixamos em cinco, o número de referências no conjunto de teste ($S = 5$). A Tabela 5.1 apresenta as taxas de acerto global para as bases BFL e IAM. Devido à regra da soma ter apresentado melhor desempenho que outras regras, as taxas apresentadas a seguir referem-se ao uso desta. Como podemos notar, aumentando o número de referências, alcançamos uma melhoria considerável utilizando poucos escritores no conjunto de treinamento. Considerando nove referências ($R = 9$), não houve melhorias significativas no desempenho, quando aumentamos o número de escritores no conjunto de treinamento. Entretanto, tais resultados referem-se ao processo de verificação, no qual temos uma comparação 1:1. Nos experimentos com a base IAM empregamos 240 escritores no conjunto de teste e fragmentos com dimensões de 256 × 128 pixels. Quanto ao número de referências, utilizamos o mesmo esquema descrito para a base BFL. Verificamos na base IAM um comportamento semelhante ao da base BFL.

Tabela 5.1: Taxa de Acerto Global (%) utilizando o descritor GLCM, variando o número de escritores no treinamento.

| | BFL | | | | IAM | | | |
|---|---|---|---|---|---|---|---|---|
| Referências | Escritores no Treinamento | | | | Escritores no Treinamento | | | |
| $R$ | 25 | 50 | 100 | 200 | 50 | 100 | 205 | 410 |
| 3 | 95,6 | 95,2 | 96,5 | 98,7 | 98,2 | 99,3 | 99,7 | 99,3 |
| 5 | 96,0 | 96,5 | 98,2 | 98,7 | 98,3 | 99,7 | 99,5 | 99,3 |
| 7 | 96,5 | 96,0 | 96,0 | 97,3 | 98,5 | 99,3 | 99,8 | 99,5 |
| 9 | 98,2 | 96,0 | 98,2 | 97,8 | 99,1 | 99,8 | 100,0 | 99,6 |

Experimentos com LBP, avaliando diferentes parâmetros, são reportados na Tabela 5.2. Tais experimentos foram avaliados com 100 escritores para treinamento e 115 para testes, tendo, $R = S = 5$. Neste primeiro experimento, fixamos o número de vizinhos

e raio ($P$ e $R$) e variamos os modelos ($LBP^{U2}_{8,2}$, $LBP^{RI}_{8,2}$ e $LBP^{RIU2}_{8,2}$). Em um segundo experimento, variamos $P$ e $R$ para um $LBP^{U2}_{x,x}$.

Tabela 5.2: Desempenho de diferentes parâmetros do LBP - base BFL.

| | Regra de Decisão | |
|---|---|---|
| LBP | Máximo | Soma |
| $LBP^{U2}_{8,2}$ | 98,70 | **99,42** |
| $LBP^{RI}_{8,2}$ | 97,39 | 96,95 |
| $LBP^{RIU2}_{8,2}$ | 95,07 | 98,55 |
| $LBP^{U2}_{16,2}$ | 98,55 | 98,41 |

Através da Tabela 5.2 é possível notar um melhor desempenho do $LBP^{U2}_{8,2}$ em relação aos outros. Aspectos como este são peculiares ao tipo de textura. Mesmo em modelos considerados mais robustos, como os invariantes à rotação, apresentaram taxas menores. Motivados com a possibilidade de melhorarmos o desempenho do sistema, aumentando o número de referências e o número de escritores, avaliamos o impacto da quantidade de referências no conjunto de treinamento $R$, fixando as referências do conjunto de teste $S$ em 5. Neste caso, o LBP padrão $LBP^{U2}_{8,2}$ foi utilizado juntamente com a regra da soma, a qual tem apresentado melhor desempenho, ilustrado na Figura 5.2.

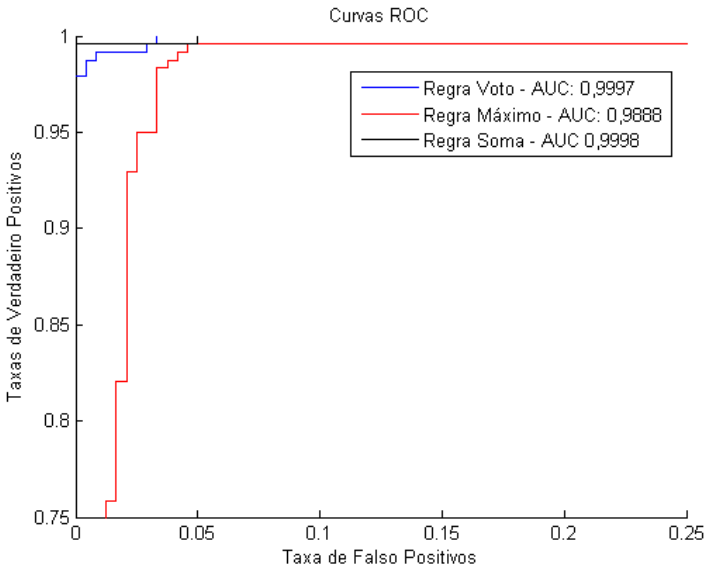



Figura 5.2: Curvas ROC para diferentes regras de decisão (Voto Majoritário, Soma e Máximo).

Na Tabela 5.3 são apresentados os resultados destes experimentos. Podemos perceber a robustez do descritor LBP comparado ao GLCM no processo de verificação de escritores. O mesmo, utilizando poucas referências e um número reduzido de escritores, alcança taxas de acertos superiores às maiores apresentadas pelo descritor GLCM. Entretanto, em relação ao objetivo do experimento, podemos notar que, aumentando o número de

referências, tem-se um impacto menor no desempenho do sistema do que se supunha. Aumentando o número de escritores foi possível um ganho entre 0,6 a 1,5 pontos percentuais. Fica claro, através da Tabela 5.3 que, com poucos escritores e um pequeno número de referências, é possível atingir ótimas taxas na verificação de escritores, tanto para base BFL quanto para a base IAM.

Tabela 5.3: Taxa de Acerto Global (%) utilizando o descritor $LBP^{U2}_{8,2}$ variando o número de escritores no treinamento.

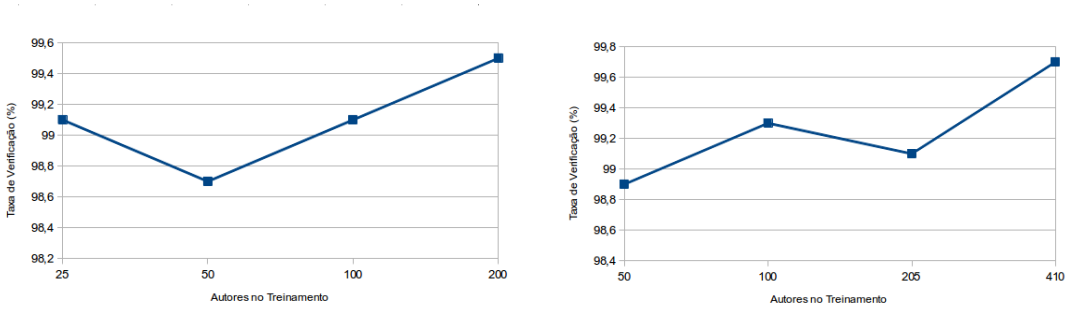
| | BFL | | | | IAM | | | |
|---|---|---|---|---|---|---|---|---|
| Referências | Escritores no Treinamento | | | | Escritores no Treinamento | | | |
| $R = S$ | 25 | 50 | 100 | 200 | 50 | 100 | 205 | 410 |
| 3 | 98,8 | 98,5 | 99,4 | 99,4 | 98,6 | 99,1 | 99,5 | 98,6 |
| 5 | 98,7 | 99,1 | 99,4 | 99,4 | 98,5 | 99,5 | 99,1 | 100,0 |
| 7 | 98,0 | 98,8 | 99,5 | 99,7 | 99,1 | 99,3 | 99,5 | 100,0 |
| 9 | 98,9 | 98,9 | 99,4 | 97,8 | 98,5 | 98,7 | 99,5 | 99,5 |

Testes combinando descritores LBP e GLCM demonstraram que o descritor LBP, isoladamente, apresenta taxas melhores que quando combinado com GLCM. Experimentos visando selecionar características do LBP, através de algoritmos genéticos, foram realizados, porém, não obtivemos sucesso. Aparentemente, para um conjunto de validação, a seleção de características conseguia uma pequena melhora, porém, ao testar as características selecionadas em um conjunto de testes, nosso desempenho caía consideravelmente.

Utilizando o descritor de textura LPQ, obtivemos taxas de acertos similares às apresentadas pelo descritor LBP. Entretanto, notamos que utilizando uma quantidade menor de escritores no conjunto de treinamento, o descritor LPQ conseguiu alcançar taxas mais altas de acerto. Outra característica interessante é o fato da base IAM ter apresentado resultados similares ou até mesmo superiores que utilizando a base BFL. Em todos os experimentos utilizamos janelas de tamanho $7 \times 7$ pixels, a qual apresentou os melhores resultados. Considerando somente a regra da soma, $R = 9$ e $S = 5$, temos os seguintes resultados para as bases BFL e IAM (Figura 5.3).

Observando a Figura 5.3, que representa a taxa de acerto global utilizando o descritor LPQ, fica evidente que, ao aumentarmos o número de escritores no conjunto de treinamento, não temos a garantia de gerarmos modelos mais robustos. Para a base BFL (Figura 5.3(a)) percebe-se que, utilizando 25 escritores no conjunto de treinamento, temos um desempenho levemente melhor que utilizando o dobro do número de escritores. O mesmo ocorre para a base IAM (Figura 5.3(b)), neste caso, empregando 100 escritores, alcançamos taxas de acertos melhores que utilizando 205 escritores. Isto corrobora com a proposta de seleção de escritores, a qual se fundamenta na ideia de que não precisamos de um grande conjunto de escritores para gerar um modelo robusto.

Devido ao tamanho dos fragmentos da base BFL ($256 \times 256$) ser maior que da IAM




(a) (b)

Figura 5.3: Desempenho das bases BFL (a) e IAM (b) utilizando o descritor LPQ com diferentes números de escritores no treinamento.

(256 × 128), aliada ao fato de termos texto-independente na base IAM, imaginávamos encontrar um desempenho mais baixo para esta base. Todavia, podemos perceber que ambas as bases apresentam taxas similares. No processo de verificação, isso é possível, pois temos comparação de 1 : 1, assim, blocos de 256 × 128 podem descrever bem o conteúdo do escritor. O fato de termos texto-independente é um problema quando escritores cedem a uma quantidade muito pequena de texto, porém, a mínima quantidade fornecida já é suficiente para gerar blocos e representar um escritor para, posteriormente, realizarmos o processo de verificação de escritor.

Analisando o tamanho do bloco de textura, podemos concluir que imagens menores têm algumas vantagens, como: menos processamento no processo de extração de textura, capacidade de gerar mais blocos de textura, consequentemente, mais referências. Hanusiak et al. [42] em seus experimentos com verificação de escritor, usando a base BFL, empregam fragmentos de tamanho 128 × 128, contudo, nenhum experimento foi realizado com o intuito de observar o impacto da dimensão dos blocos de textura. Motivados com isso, avaliamos diferentes dimensões dos blocos de textura. O desempenho apresentado na Tabela 5.4, refere-se à base BFL, com 100 escritores no conjunto de treinamento e 115 para testes, ($R = S = 5$). O LBP utilizado foi o $LBP^{U2}_{8,2}$ e LPQ com janelas 7 × 7, em ambos os casos, utilizamos a regra da soma.

De maneira geral, avaliando a Tabela 5.4, percebemos que ao utilizar fragmentos de 256 pixels de altura conseguimos taxas melhores, no entanto, isto não é possível para base IAM. Outra observação é que fragmentos de tamanho reduzido, como 64 × 64, não são convenientes, isto devido à pouca informação relevante presente na imagem. Por outro lado, fragmentos muito grandes (768 pixels de largura) também não são adequados. Neste caso, é provável que exista uma variação intrapessoal na escrita do mesmo e, quanto maior o fragmento, maior o desvio de comportamento existente. É importante salientar que devemos levar em conta o descritor utilizado, já que utilizando LPQ, não houve melhoria em blocos com dimensões maiores que 256 pixels. Desta forma, em nossos experimentos,

Tabela 5.4: Desempenho do LBP e LPQ na verificação de escritor, considerando diferentes tamanhos de fragmentos - base BFL.

| Tamanho (Larg. $\times Alt.$) | Número de Fragmentos | Taxa de Acerto (%) | |
|---|---|---|---|
| | | LBP | LPQ |
| 64 × 64 | 114 | 87,4 | 94,4 |
| 128 × 128 | 57 | 98,2 | 98,7 |
| 209 × 128 | 33 | 97,8 | 98,7 |
| 256 × 128 | 27 | 98,7 | 97,4 |
| 329 × 128 | 21 | 99,5 | 97,9 |
| 460 × 128 | 15 | 98,2 | 97,9 |
| 768 × 128 | 9 | 97,3 | 98,3 |
| 209 × 256 | 33 | 98,2 | 97,0 |
| 256 × 256 | 27 | 98,7 | 98,7 |
| 309 × 256 | 21 | 99,5 | 98,7 |
| 460 × 256 | 15 | 99,1 | 98,3 |
| 768 × 256 | 9 | 99,1 | 98,7 |

padronizamos em 256 × 256 pixels os fragmentos das bases BFL e *Firemaker*.

Devemos lembrar que, ao utilizar fragmentos maiores, diminuímos o número de amostras. Entretanto, como utilizamos, no máximo, nove referências para o processo de treinamento/teste, 768 pixels de largura é a máxima dimensão possível para conseguirmos o número de amostras necessárias. Todavia, tais dimensões não puderam ser avaliadas para base IAM, pois ao utilizar fragmentos de 256 × 128, conseguimos entre 9 e 18 fragmentos (conforme a quantidade de texto cedido pelo escritor). Desta forma, caso utilizássemos tamanhos excedentes ao descrito, seria impossível gerar nove fragmentos para utilizarmos na abordagem proposta. Desta forma, experimentos realizados com a base IAM terão blocos de tamanho 256 × 128 pixels.

A Tabela 5.5 reporta os resultados obtidos por diferentes autores, utilizando a base BFL com a mesma quantidade de escritores no conjunto de teste (115). Podemos perceber que, no processo de verificação de escritor, as taxas alcançadas são superiores às apresentadas na literatura. O mesmo ocorre para base IAM, na qual o desempenho apresentado é similar. Nesta tabela, apresentamos as taxas empregando o número máximo de escritores no conjunto de treinamento com o $R = 9$ e $S = 5$.

Através dos experimentos realizados, foi possível perceber que, mesmo utilizando um pequeno número de escritores no conjunto de treinamento, é possível construir um modelo robusto para o processo de verificação de escritor. A avaliação da quantidade de referências usada no conjunto de treinamento foi essencial para percebermos que o número de referência é menos importante do que se supunha para o processo de verificação de escritor. Percebemos, também, a robustez dos descritores LBP e LPQ para esta aplicação, entretanto, nosso grande desafio é atingir um excelente desempenho na identificação de

Tabela 5.5: Comparação entre GLCM, LBP e LPQ para as bases BFL e IAM.

| Descritor | Taxa de Acerto (%) | |
|---|---|---|
| | Base BFL | BaseIAM |
| GLCM (Entropia) [42] | 95,9 | - |
| GLCM (4 características) [61] | 94,9 | - |
| GLCM (Abordagem empregada) | 97,8 | 99,6 |
| LBP (Abordagem empregada) | 97,8 | 99,5 |
| LPQ (Abordagem empregada) | 99,5 | 99,7 |

escritores, já que os experimentos com verificação de escritor fazem parte desta tese para demonstrar a robustez da abordagem. Assim, podemos concluir que a abordagem escritor-independente apresenta-se como uma abordagem viável, utilizando um pequeno número de escritores para o conjunto de treinamento e poucas amostras de referências. O restante deste trabalho foca, unicamente, no processo de identificação de escritores.

### 5.1.2 Identificação de Escritores

O problema de identificação de escritores consiste em determinar a identidade $I$ entre todos os escritores envolvidos no processo. Utilizando a representação da dissimilaridade, o processo de identificação pode ser realizado chamando o processo de verificação $n$ vezes, onde $n$ é o número de escritores existentes no conjunto de teste. Assim, o processo de identificação é dado como correto quando o escritor procurado possuir a maior estimação de probabilidade *a posteriori*. No entanto, o sistema de identificação também pode fornecer uma lista de documentos que são semelhantes ao documento consultado. O tamanho desta lista, também conhecida como *hit list*, pode variar, por exemplo, 1, 5, ou 10. Desta forma, em alguns casos, apresentaremos os resultados em função do Top-1, Top-5 e Top-10 para o desempenho de identificação de escritor. Apresentaremos, a seguir, os resultados obtidos na identificação de escritores utilizando as bases BFL e IAM. Apresentaremos nossos experimentos, seguindo a mesma linha de raciocínio empregada na Seção 5.1.1.

Experimentos passados demonstraram que o descritor GLCM é robusto na verificação de escritores, porém, para identificação, o mesmo apresentou baixas taxas de acerto. Experimentos realizados com LBP demonstraram desempenho superior em relação ao GLCM. A Tabela 5.6 demonstra resultados para Top-1 alcançados com GLCM, LBP e LPQ. Empregamos a base BFL utilizando as melhores combinações apresentadas no processo de verificação: Fragmentos de dimensões $256 \times 256$, $R = S = 5$, descritores $LBP^{U2}_{8,2}$, GLCM (Energia, Contraste, Correlação e Homogeneidade com distâncias concatenadas 1 à 5), LPQ com janela $7 \times 7$ e, a Regra da Soma foi utilizado o método de combinação das saídas dos classificadores.

Nota-se que o descritor GLCM é comparável ao LBP no processo de verificação, en-

Tabela 5.6: Taxa de Acerto Global (%) usando descritores GLCM, LBP e LPQ - base BFL.

| | Escritores no Treinamento | | | |
|---|---|---|---|---|
| Descritor | 25 | 50 | 100 | 200 |
| GLCM | 47,8 | 45,2 | 27,8 | 46,0 |
| LBP | 69,5 | 73,0 | 82,6 | 86,9 |
| LPQ | 88,7 | 87,0 | 92,2 | 94,0 |

tretanto, na identificação apresentou baixas taxas, deixando clara a robustez do descritor LBP. Desta maneira, nossos experimentos com GLCM encerram-se aqui. Dando sequência aos nossos experimentos, dos quais continuaremos com os casos de sucesso obtidos, avaliamos o impacto da quantidade de escritores presentes nos conjuntos de treinamento em função das diferentes regras de decisão. Adotamos, até o momento, a regra da soma em todos os nossos experimentos, entretanto, não avaliamos o impacto da mesma no processo de identificação de escritores. A Tabela 5.7, demonstra as taxas de acerto, variando o número de escritores no conjunto de treinamento para as quatro principais regras de fusão. Fixamos o número de referências do conjunto de treinamento e teste em cinco referências ($R = S = 5$). A Tabela 5.7 demonstra as taxas alcançadas com descritor LPQ, contudo, experimentos com descritor LBP apresentam resultados similares, percebendo a robustez do descritor LPQ ao possuirmos poucos escritores no conjunto de treinamento.

Tabela 5.7: Taxa de Acerto Global (%) usando LPQ e $R = S = 5$ - base BFL e IAM.

| | BFL | | | | IAM | | | |
|---|---|---|---|---|---|---|---|---|
| Regra de | Escritores no Treinamento | | | | Escritores no Treinamento | | | |
| Fusão | 25 | 50 | 100 | 200 | 50 | 100 | 205 | 410 |
| Soma | 88,7 | 87,0 | 92,2 | 94,0 | 74,6 | 79,2 | 80,5 | 82,5 |
| Máximo | 80,0 | 93,1 | 94,7 | 90,5 | 55,9 | 70,0 | 78,4 | 79,6 |
| Produto | 87,9 | 84,4 | 89,6 | 92,2 | 72,1 | 78,0 | 78,4 | 80,0 |
| Mediana | 87,0 | 91,3 | **94,8** | 93,1 | 73,8 | 81,7 | 84,6 | **88,0** |

Através da Tabela 5.7 é possível perceber que a melhora de desempenho do sistema, aumentando o número de escritores no conjunto de treinamento, não é uma regra. Em muitos dos casos, o ganho ao dobrar o número de escritores do conjunto de treinamento não contribuiu para melhora no sistema. Isso mostra, novamente, que a seleção de escritores pode ser benéfica para o processo. Percebe-se que as melhores taxas de acertos alcançadas no processo de identificação de escritores são bem menores que as alcançadas através do processo de verificação de escritor. Um detalhe quanto às regras de decisão, é que a regra da Mediana apresentou bons resultados utilizando um grande número de escritores no treinamento. Utilizando poucos escritores, a regra não apresentou bom desempenho.

Mesmo não apresentando os melhores resultados, podemos notar que a regra da soma é bastante estável e sempre apresenta boas taxas.

A Tabela 5.7 nos mostra que, em alguns casos, aumentando o número de escritores no conjunto de treinamento, podemos melhorar o desempenho do sistema. Utilizando a base BFL, percebemos que, em alguns casos, ao aumentar o número de escritores, reduzimos o desempenho do sistema. Experimentos com a base IAM demonstraram que, aumentando o número de escritores é possível melhorar as taxas de acerto. Entretanto, nos experimentos nos quais dobramos o número de escritores de 205 para 410, alcançamos, no melhor caso, um aumento de 3,6 pontos percentuais, uma taxa relativamente baixa pela quantidade de escritores incluídos no conjunto de treinamento. Percebemos, também, que o descritor LPQ pode ser visto como mais robusto ao utilizarmos um número reduzido de escritores no treinamento. A melhor taxa alcançada com a base BFL foi de 94,8% de acerto na identificação, utilizando 100 escritores no treinamento. Para base IAM, alcançamos 88,0% de identificação, empregando o número máximo de escritores (410). Utilizando o descritor LBP, as melhores taxas alcançadas foram empregando a regra da Mediana, usando o número máximo de escritores, reportando taxas de acertos globais de 87,2% e 84,5 para a BFL e IAM, respectivamente.

Embasados nos experimentos anteriores, resolvemos investigar o impacto ao possuirmos diferentes números de referências no conjunto de treinamento $R = 3, 5, 7$ e 9, fixando em cinco, o número de referências para o conjunto de teste ($S = 5$). Neste caso, utilizamos o descritor LPQ e a regra da Mediana para combinar as saídas do classificador. Podemos verificar estes desempenhos na Tabela 5.8.

Tabela 5.8: Taxa de Acerto Global (%) para diferentes números de escritores e de referências no treinamento ($R$) - base BFL e IAM.

| Número de Referências | BFL Escritores no Treinamento | | | | IAM Escritores no Treinamento | | | |
|---|---|---|---|---|---|---|---|---|
| $R$ | 25 | 50 | 100 | 200 | 50 | 100 | 205 | 410 |
| 3 | 88,7 | 90,5 | 94,8 | 94,0 | 85,9 | 82,5 | 83,0 | 86,3 |
| 5 | 94,7 | 91,3 | 94,8 | 93,1 | 73,8 | 81,7 | 84,6 | 88,0 |
| 7 | 92,2 | 92,2 | 92,2 | 93,1 | 79,6 | 80,5 | 86,7 | 88,4 |
| 9 | **95,7** | 92,2 | 93,5 | 93,5 | 87,5 | 88,0 | 86,3 | **89,6** |

Através destes experimentos, podemos concluir que, aumentando o número de referências ($R$) no conjunto de treinamento, podemos melhorar as taxas de acertos. Foi possível perceber também que o descritor LPQ apresenta boas taxas, mesmo utilizando poucas referências, deixando clara a robustez do descritor LPQ. Entretanto, experimentos em ambas as bases nos revelam que, nem sempre, aumentando o número de escritores no conjunto de treinamento, conseguimos uma melhora no desempenho. Isto corrobora com nossa tese de que não necessitamos de um grande número de escritores no treinamento,

mas sim, de escritores que, combinados, gerem um modelo de classificação robusto. Resultados com LBP apresentaram taxas próximas ao utilizarmos um grande número de escritores, no entanto, utilizando um conjunto de treinamento com poucos escritores, o LPQ demonstra-se mais robusto. Na base BFL, a melhor taxa foi 92,1% e, para IAM, alcançamos 88,0% usando o descritor LBP.

É importante lembrar que nove é o número máximo de referências que podemos encontrar para os 650 escritores da base IAM, isto justifica tal valor. Percebe-se, então, que para ambas as bases, alcançamos uma redução nas taxas de erro, aumentando o número de referências $R$ de três para nove.

Observando a evolução que alcançamos, aumentando o número de referências no conjunto de treinamento ($R$), nos propomos investigar o impacto do número de referências no conjunto de teste ($S$). A ideia é que, ao contar com mais informações para a tomada de decisões, poderíamos alcançar um melhor desempenho. Assim, as Tabelas 5.9 e 5.10 demonstram a evolução do número de referências no teste ($S$) para as bases BFL e IAM, respectivamente. Em ambos os casos, temos usado o número máximo de escritores disponível para treinamento (200 e 410), fixando $R$ em nove referências. Os resultados apresentados quanto ao Top-5 e Top-10 referem-se à regra da Mediana.

Tabela 5.9: Avaliação do número de referências ($S$) na identificação de escritor - base BFL

| | LBP | | | | LPQ | | | |
|---|---|---|---|---|---|---|---|---|
| Regra de | Número de Referências | | | | Número de Referências | | | |
| Fusão | S=3 | S=5 | S=7 | S=9 | S=3 | S=5 | S=7 | S=9 |
| Soma | 78,2 | 92,2 | 96,5 | 93,9 | 88,7 | 92,2 | 99,2 | 99,2 |
| Máximo | 71,3 | 74,8 | 71,3 | 73,0 | 89,6 | 92,2 | 94,8 | 93,1 |
| Produto | 77,4 | 89,5 | 93,9 | 90,4 | 88,7 | 90,5 | 95,7 | 96,6 |
| Mediana | 76,5 | 90,4 | 94,6 | **94,7** | 86,1 | 90,5 | 97,4 | **99,2** |
| Top-5 | 97,3 | 99,2 | 99,2 | 99,2 | 95,7 | 96,6 | 100 | 99,2 |
| Top-10 | 99,2 | 99,2 | 99,2 | 99,2 | 98,3 | 99,9 | 100 | 100 |

Ao adicionar blocos de textura no conjunto de teste, foi possível melhorar consideravelmente o desempenho do sistema. Em ambas as bases de dados, o descritor LPQ apresentou as melhores taxas, 99,2% para BFL e 96,7% para base IAM. Ao analisar os erros em relação ao Top-5 e Top-10, notamos que, na maioria dos casos, a classe correta não estava muito longe da classe eleita. Percebemos ainda que, mesmo utilizando um pequeno número de escritores no conjunto de treinamento, ao avaliarmos um Top-5 ou Top-10, conseguimos uma excelente taxa de acertos.

É comum o uso de curvas CMC (*Cumulative Match Characteristic*) [15] para avaliação do desempenho de sistemas biométricos, que têm como função a identificação ou reconhecimento ($1 : N$). O propósito é avaliar a probabilidade de o sistema retornar um candidato em uma lista de tamanho $N$. Esta curva demonstra a taxa de identificação

Tabela 5.10: Avaliação do número de referências ($S$) na identificação de escritor - base IAM

| Regra de Fusão | LBP Número de Referências | | | | LPQ Número de Referências | | | |
|---|---|---|---|---|---|---|---|---|
| | S=3 | S=5 | S=7 | S=9 | S=3 | S=5 | S=7 | S=9 |
| Soma | 69,2 | 83,3 | 90,4 | 92,9 | 68,8 | 83,8 | 91,3 | 90,9 |
| Máximo | 72,5 | 69,2 | 75,8 | 77,5 | 76,3 | 80,5 | 85,5 | 87,1 |
| Produto | 68,7 | 80,2 | 87,5 | 90,4 | 68,4 | 80,9 | 87,5 | 89,6 |
| Mediana | 68,7 | 83,7 | 88,3 | **94,6** | 68,0 | 89,6 | 93,8 | **96,7** |
| Top-5 | 94,1 | 98,3 | 100 | 99,6 | 91,3 | 98,4 | 99,2 | 100 |
| Top-10 | 97,9 | 99,2 | 100 | 100 | 97,5 | 98,8 | 99,2 | 100 |

correta em função de um *rank*. A Figura 5.4 apresenta as curvas CMC, utilizando a regra da Mediana.

Diversos trabalhos utilizando estas bases reportam taxas de acerto próximas a 100% para o processo de identificação, entretanto, muitas vezes utilizam um subconjunto muito pequeno da base de dados (*vide* Tabela 3.2), tornando difícil uma real comparação.

É importante destacar que este tipo de aplicação tem por finalidade auxiliar um perito forense nas suas investigações. Desta forma, o perito não necessita de um sistema que apresente taxas de 100% para Top-1, mas de um sistema que apresente uma lista reduzida, a qual exista uma alta probabilidade da resposta correta estar contida nesta lista. Podemos notar, a partir das Curvas CMC, que conseguimos alcançar desempenho acima de 99% em ambas as bases de dados, considerando Top-5.

### 5.1.3 Abordagem Escritor-Dependente × Escritor-Independente

Analisando os resultados apresentados, podemos nos questionar a qual método se deve o bom desempenho deste trabalho. A princípio, podemos concluir que a união dos métodos pode ter sido responsável pelo bom desempenho do sistema. Entretanto, isto se deve, principalmente: (*i*) ao método de geração de textura, juntamente com os descritores empregados (*ii*) à abordagem de dissimilaridade, ou (*iii*) à combinação da dissimilaridade com características de textura?

Visando entender melhor estes detalhes, utilizamos duas outras abordagens de escritor-dependente neste trabalho. Em geral, estas abordagens apresentam bons resultados, porém, sua maior desvantagem é que, para cada novo escritor, um novo modelo deve ser construído. Outra questão importante nesta estratégia é a necessidade de uma quantidade de dados suficiente para treinar um modelo confiável. No nosso caso, o número de amostras disponíveis para o aprendizado é pequena, 9 blocos de textura por escritor.

O primeiro modelo escritor-dependente implementado foi um SVM multiclasse, usando uma abordagem de comparação em pares. Nesta estratégia, o número de classificadores a

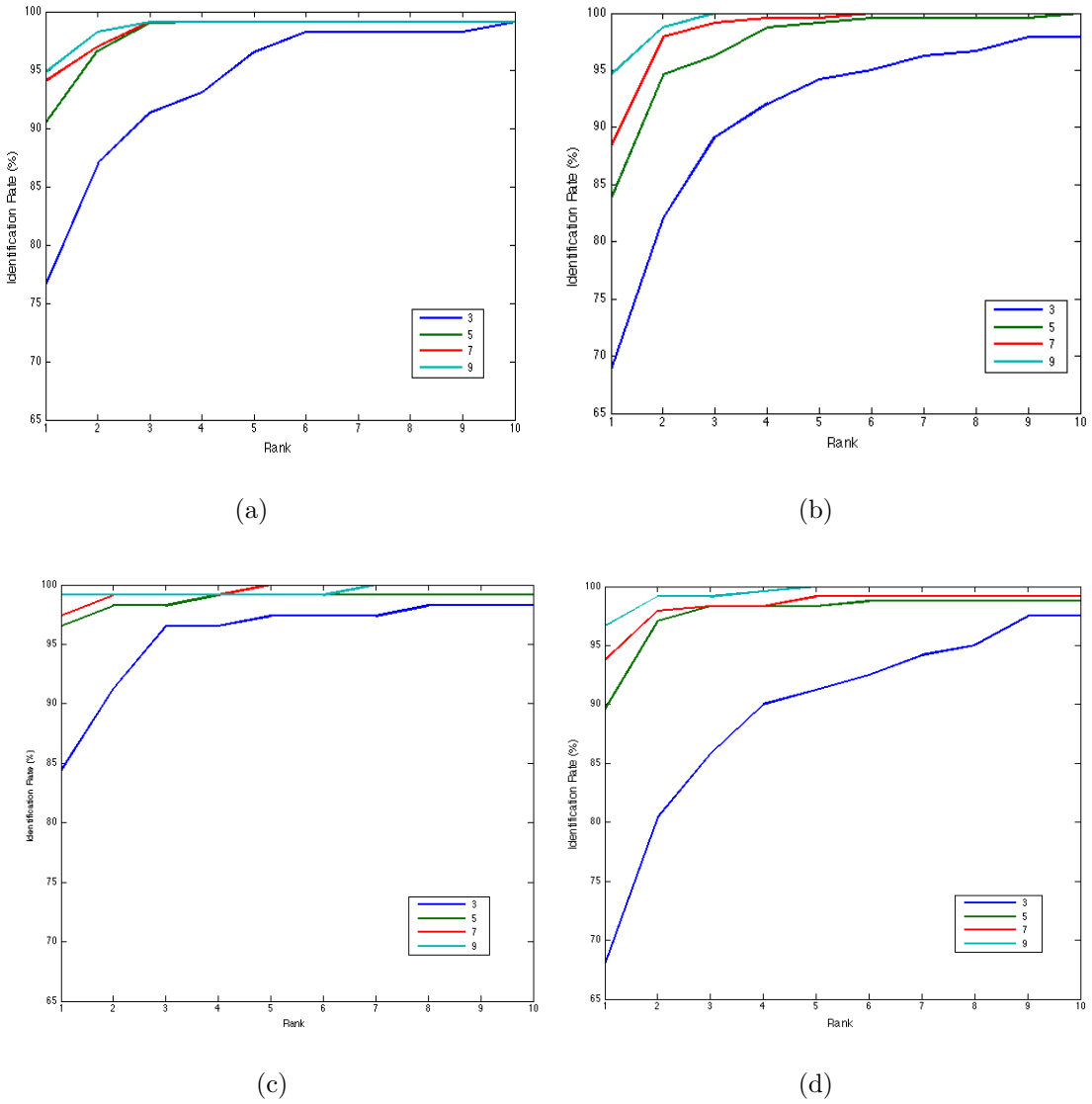



Figura 5.4: Curvas CMC: Figura (a) Descritor LBP - base BFL; Figura (b) Descritor LPB - base IAM; Figura (c) Descritor LPQ - base BFL; Figura (d) Descritor LPQ - base IAM.

serem formados é dado por $q(q-1)/2$, onde $q$ é o número de classes (escritores, no nosso caso).

A segunda estratégia é a decomposição de um-contra-todos, que trabalha na construção de um SVM $\omega_i$ para cada classe $q$, com o objetivo de separar esta classe de todas as outras. Comparando as duas abordagens dependentes de escritor, a estratégia um-contra-todos é mais adequada para esta aplicação, já que um novo modelo deve ser treinado cada vez que um novo escritor é inserido no sistema.

A fim de manter o mesmo protocolo, o número de classes ($q$) utilizado nestes experimentos é o mesmo dos experimentos anteriores, ou seja, 115 para base BFL e 240 para base IAM. Ao contrário do protocolo da abordagem de escritor-independente empregada até o momento, nas abordagens dependentes de escritores, precisamos de amostras do

escritor nos conjuntos de treinamento e teste. No caso da base BFL, todos os escritores possuem três cartas manuscritas, assim, iremos utilizar duas cartas para o treinamento (18 blocos de textura) e uma para teste (9 blocos de textura), todas com dimensões de $256 \times 256$ pixels. Na base IAM, como o número de cartas cedidas por cada escritor varia bastante, selecionamos 240 escritores que possuíam duas cartas, utilizando uma para treinamento e a segunda para testes. Desta forma, conseguimos nove imagens de textura para treinamento e nove para testes, com dimensões de $256 \times 128$ pixels. A Tabela 5.11 apresenta os resultados alcançados em cada abordagem.

Tabela 5.11: Taxa de Acerto Global (%) das diferentes estratégias de classificação usando descritor LPQ.

| Estratégia | BFL | IAM |
|---|---|---|
| Dissimilaridade | 99,2 | 96,7 |
| Comparação em pares | 98,2 | 85,2 |
| Um-contra-todos | 97,4 | 88,3 |

Em relação às questões levantadas no início desta seção, estes resultados demonstram que a abordagem baseada na dissimilaridade apresenta desempenho superior às demais. No caso da base BFL, para a qual existem mais amostras disponíveis, temos taxas mais altas; contudo, a abordagem empregada nesta tese, apresenta ligeira vantagem. Uma diferença considerável pode ser observada na base IAM. Isto, devido ao conjunto de treinamento ser menor (1 carta para treinamento e 1 carta para teste). Desta maneira, podemos concluir que a abordagem de escritor-independente em conjunto com a dissimilaridade, utilizando a representação dos escritores através de um processo de geração de textura unidos a descritores robustos, constrói uma abordagem adequada, mesmo quando apenas poucas amostras por escritor estão disponíveis.

## 5.2 Avaliação de Diferentes Estilos de Escrita

O principal objetivo desta seção é demonstrar que a abordagem empregada para identificação de escritores apresentada anteriormente (*vide* seção 5.1.2) é um método robusto, independente do estilo de escrita utilizado. Dois estilos de escritas foram avaliados na seção anterior, o estilo texto-dependente e texto-independente. Nesta, percebemos um melhor desempenho utilizando texto-dependente. Entretanto, devido à quantidade de escritores nos conjuntos de treinamento e testes serem diferentes, é difícil mensurar a real diferença entre ambas. Desta forma, utilizaremos, nestes experimentos, a base *Firemaker*, a qual possui os estilos de escrita já utilizados anteriormente, e mais dois outros estilos: o estilo caixa alta e a falsificação.

Nestes experimentos, utilizamos 150 escritores para os conjuntos de teste. Empregamos este número a fim de comparar os resultados obtidos com os apresentados por

Schomaker et al. [87], o qual emprega a mesma quantidade de escritores. Ao utilizar 150 escritores no conjunto de teste, sobraram 100 escritores para o conjunto de treinamento. Desta forma, avaliaremos o impacto de possuir mais e menos escritores no conjunto de treinamento, sendo uma partição com todos os escritores restantes, 100; e outra, utilizando um número reduzido de escritores, 20. Observando os resultados alcançados com as bases BFL e IAM empregamos, nestes experimentos, descritores e parâmetros que caracterizaram sucesso. Assim, utilizamos os descritores de textura $LBP^{U2}_{8,2}$ e LPQ $7 \times 7$. Em todos os experimentos realizados com diferentes estilos de escrita consideramos nove referências para o treinamento e nove para o teste, $R = S = 9$, isto devido à observação dos resultados anteriores demonstrarem bom desempenho ao empregar esta configuração. Avaliamos diferentes regras de fusão, entretanto, a regra da soma apresentou-se mais estável para poucos e muitos escritores, diferente da regra da mediana, que apresentava bons resultados, unicamente, quando tínhamos muitos escritores.

Inicialmente, avaliamos o desempenho para os diferentes estilos de escrita na identificação de escritor. Assim, empregamos o mesmo estilo de escrita para treinamento e teste. Os autores do trabalho que descrevem a base *Firemaker* [88] utilizam os termos *Cópia* e *Natural* para descrever os estilos, texto-dependente e texto-independente, respectivamente. Neste trabalho, adotaremos tais nomenclaturas. Avaliando o desempenho dos três diferentes estilos de escrita, verificamos que o melhor resultado foi reportado usando texto-dependente (*Cópia*). Tal desempenho pode ser justificado, pois todos os escritores possuem a mesma quantidade de texto. Em tese, podemos supor que em ambos os estilos, *Cópia* e *Natural*, a escrita é similar (Figura 5.5). Contudo, a principal diferença entre estes estilos são o número de linhas e/ou palavras escritas o que, em geral, no estilo texto-independente varia de 35% a 60% do conteúdo presente no estilo texto-dependente. Observando a Tabela 5.12 podemos notar que o estilo *Natural* apresentou bons resultados, próximos aos apresentados no estilo *Cópia*. Analisando os blocos de textura presentes na Figura 5.5, podemos perceber que o conteúdo textural presente no bloco gerado pelo estilo *Cópia* é maior que o gerado pelo estilo *Natural*, sendo esta a principal justificativa da diferença de desempenho.

Em síntese, o estilo caixa alta, apresentou as mais baixas taxas dentre os três estilos, entretanto, para o melhor caso, apresentando taxas de acerto de 93%, é bem superior ao reportado por Schomaker et al. [87] de 70%. Neste caso, devido ao escritor empregar um estilo de escrita, para a qual pode não estar habituado, podemos considerar a existência de uma maior variabilidade intraclasse.

Através da Tabela 5.12 podemos notar uma maior robustez do descritor LPQ, principalmente ao utilizar um pequeno número de escritores para treinamento. Outra característica que pode ser notada facilmente é o fato de que, reduzindo-se o número de escritores de 100 para 20, alcançamos taxas de acerto iguais ou melhores.

Analisando o desempenho em relação à quantidade de escritores no conjunto de treinamento, é fácil notar que, ao aumentar o número de escritores no conjunto de treinamento,

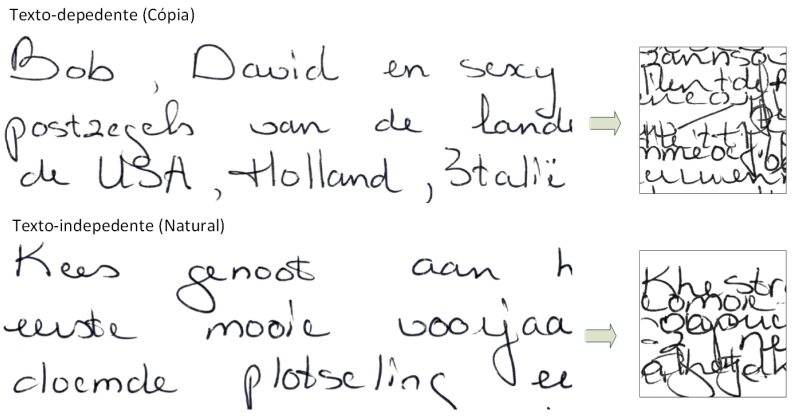



Figura 5.5: Similaridade entre escrita no estilo texto-dependente e texto-independente.

Tabela 5.12: Taxa de Acerto Global (%) avaliando diferentes estilos de escrita.

| Estilo da Escrita | | Escritores Treino 100 | | Escritores Treino 20 | |
|---|---|---|---|---|---|
| Treinamento | Teste | LBP | LPQ | LBP | LPQ |
| *Cópia* | *Cópia* | 98,0 | 98,0 | 96,0 | **98,0** |
| *Natural* | *Natural* | 91,0 | 94,0 | 79,0 | 91,0 |
| *Caixa alta* | *Caixa alta* | 93,0 | 93,0 | 87,0 | 89,0 |
| *Mix* | *Cópia* | 90,0 | 95,0 | 95,0 | 94,0 |
| *Mix* | *Natural* | 79,0 | 87,0 | 77,0 | 86,0 |
| *Mix* | *Caixa Alta* | 76,0 | 78,0 | 74,0 | 80,0 |

podemos melhorar as taxas de acerto, entretanto, essa melhora fica mais clara utilizando o descritor LBP. No caso do estilo *Cópia*, o qual apresentou as melhores taxas, conseguimos 98% de acerto em ambos os conjuntos de treinamento (20 e 100 escritores). Isto reforça a tese de que não necessitamos de um grande conjunto de escritores no treinamento. Assim, como nos experimentos com outras bases, o descritor LPQ mantém-se robusto, mesmo utilizando poucos escritores no treinamento.

Comparando os resultados alcançados com a base *Firemaker*, reportado por Schomaker et al. [87], podemos notar que as taxas para o estilo *Cópia* são muito próximas (97%). Entretanto, para os estilos *Caixa Alta* e *Natural*, cuja variabilidade na escrita é mais proeminente, conseguimos um melhora de até 24 pontos percentuais em relação ao trabalho de Schomaker et al., o qual descreve taxas de 70% para identificação em ambos os estilos. Isto corrobora com o argumento de que a textura é uma boa alternativa para identificação de escritores. As curvas ROC, apresentadas na Figura 5.6, demonstram que utilizando o estilo *Cópia* foi possível alcançar os melhores resultados no processo de identificação de escritores. Um dos fatores que deve ser levado em conta para o estilo *Caixa alta* é que o conteúdo escrito foi baseado em um texto de referência, desta forma, todos os escritores têm a mesma quantidade de texto. Os resultados alcançados com a base IAM, a qual é texto-independente, é próximo aos apresentados utilizando a base *Firemaker* estilo *Natu-*

*ral*, em ambas temos as mesmas limitações devido à alta variação da quantidade de texto manuscrito cedido pelo escritor.

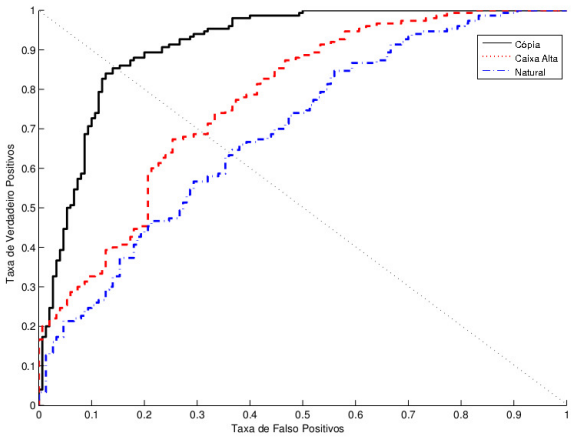



Figura 5.6: Curvas ROC produzidas através do descritor LPQ - (Treinamento = 150, Teste = 100).

A segunda parte da Tabela 5.12 nos apresenta taxas de acerto utilizando uma mistura dos estilos de escrita, no conjunto de treinamento. Como descrito, os experimentos utilizando a base *Firemaker* empregam nove blocos de textura ($R = 9$). Para isso, utilizaram-se três blocos do estilo *Cópia*, três de estilo *Natural* e três *Caixa Alta*. Tecnicamente, os três estilos contêm características únicas dos escritores, sendo possível que um modelo gerado através desta mistura de estilos possa ser genérico bastante para produzir resultados melhores que utilizando um único estilo de escrita. A segunda parte da Tabela 5.12 nos mostra que isso não é verdade. O desempenho utilizando o mesmo estilo no conjunto de treinamento e teste apresenta um melhor desempenho. Para o estilo *Cópia*, alcançamos taxas bem aceitáveis, como 94%, entretanto, para o estilo *Natural*, a melhor taxa reportada foi de 80%. Através da mistura de estilos (*Mix*) não conseguimos melhorar o desempenho do sistema; com isso, podemos notar mais uma vez a robustez do descritor de textura para este tipo de aplicação. Não encontramos em literatura experimentos similares, assim, não foi possível comparar o desempenho descrito com outros trabalhos.

Por fim, a base *Firemaker* possui um estilo descrito como *Falsificação*. Este estilo é gerado de forma que o próprio escritor é forçado a distorcer sua escrita, gerando este estilo. Nosso objetivo é demonstrar que, mesmo alterando seu estilo de escrita manuscrita, é possível identificar um escritor, pois ao utilizar a textura como característica, é possível obter detalhes que o escritor não consegue disfarçar como a pressão e outras informações únicas da grafia do escritor. É possível que o escritor consiga, por algum tempo, distorcer bem sua escrita, porém, ao longo do texto, tende a deixar informações da sua escrita pessoal.

A Tabela 5.13 apresenta o desempenho alcançado através do estilo *Falsificação*. Pode-

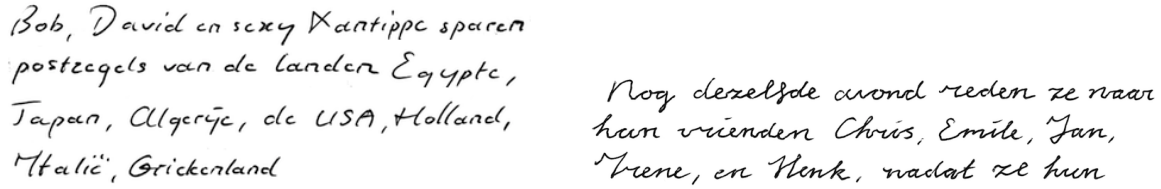
mos facilmente perceber um comportamento contrário em relação ao desempenho e estilo de escrita empregado no conjunto de treinamento. Nestes experimentos, o estilo *Natural* foi quem apresentou o melhor desempenho (94%) e o estilo *Cópia* apresentou os piores resultados. É provável que esta inversão se deva ao fato de possuirmos uma melhor representabilidade da variação de escrita do escritor no estilo *Natural*, pois no estilo *Cópia*, temos uma maior quantidade de texto e um padrão mais comportado, gerando modelos mais especialistas e menos genéricos. Entretanto, para o processo de identificar uma falsificação, amostras com maior variabilidade de escrita no conjunto de treinamento podem promover melhor desempenho. Podemos comprovar esta hipótese observando as taxas da segunda parte da Tabela 5.13, na qual podemos encontrar o desempenho utilizando uma mistura dos três estilos. Percebe-se que, como nos experimentos anteriores, não obtivemos melhorias utilizando o conjunto *Mix*. Assim, a mistura dos três estilos de escrita apresentou taxas melhores que empregando o estilo *Cópia*. Desta forma, podemos entender que o estilo *Natural* contém uma variabilidade boa para identificar falsificações. Contudo, mesmo neste cenário, as taxas de erro para identificação de falsificação são aceitáveis.

Tabela 5.13: Taxa de Acerto Global (%) das diferentes estratégias de classificação empregando descritor LPQ.

| Estilo da Escrita | | Escritores Treino 100 | | Escritores Treino 20 | |
|---|---|---|---|---|---|
| Treinamento | Teste | LBP | LPQ | LBP | LPQ |
| *Cópia* | *Falsificação* | 78,0 | 84,0 | 78,0 | 78,0 |
| *Natural* | *Falsificação* | 91,0 | 91,0 | 84,0 | 86,0 |
| *Caixa Alta* | *Falsificação* | 92,0 | **94,0** | 88,0 | 90,0 |
| *Mix* | *Falsificação* | 84,0 | 86,0 | 76,0 | 72,0 |

Como nos experimentos anteriores, o LPQ apresentou o melhor desempenho e, utilizando 100 escritores no conjunto de treinamento, conseguimos taxas relativamente melhores. Schomaker et al. [87], utilizando um esquema similar em que considera vários estilos para treinar um modelo, reporta taxa de acerto de 50%. A Figura 5.7 compara o estilo de escrita *Natural* Figura 5.7(a) e do texto falsificado pelo próprio escritor (Figura 5.7(b)). Ele demonstra o quão difícil pode ser a tarefa de identificar um texto manuscrito utilizando o estilo *Falsificação*.

Os resultados alcançados com a base *Firemaker* são similares aos apresentados utilizando as bases BFL e IAM. Através da base de texto-dependente (BFL) conseguimos uma taxa de 99,2%, utilizando 115 escritores no conjunto de teste, e 200, no treinamento. A base *Firemaker* com texto-dependente (*Cópia*) apresentou taxas de 98% de acerto. Já para a base IAM, com texto-independente, alcançamos 96,7% de acerto, empregando 240 escritores no conjunto de testes e 410 para treinamento. Comparando a base *Firemaker*



(a) (b)

Figura 5.7: Dois blocos de texto do mesmo escritor : Figura (a) *Natural* e Figura (b) *Falsificação*.

com texto-independente (*Natural*), obtivemos taxas de 94% de acerto na identificação. Devemos lembrar que, para ambos os casos, com a base *Firemaker* utilizamos 150 escritores no conjunto de teste e 100 para treinamento. Resumindo, esta seção demonstra o quão robusto é o conjunto de técnicas empregadas no processo de identificação de escritores, demonstrando sua eficiência para diferentes estilos de escrita e, principalmente, a *Falsificação*. Detalhes sobre estes experimentos estão disponíveis em [13].

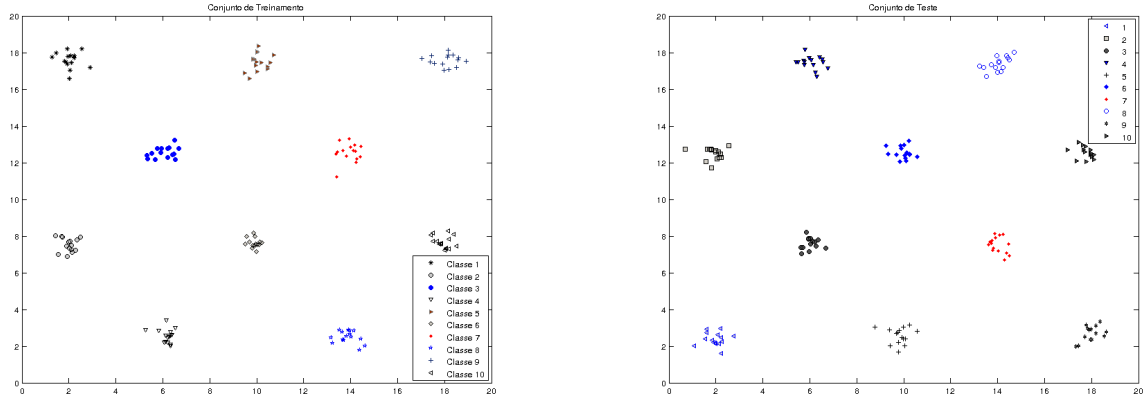
## 5.3 Seleção de Escritores

Embasados nos experimentos anteriores e, seguindo as diretivas desta pesquisa, apresentaremos, nesta seção, uma abordagem na qual selecionaremos escritores para compor o conjunto de treinamento. Uma proposta similar é apresentada por Garcia et al. [37] para seleção de protótipos. Discutiremos a proposta de seleção de escritores em duas etapas, na primeira etapa, utilizaremos uma base sintética como prova de conceito demonstrando a abordagem de seleção de escritores. A segunda etapa apresenta experimentos e resultados da proposta de seleção de escritores para as bases BFL, IAM e *Firemaker*.

### 5.3.1 Base Sintética - Prova de Conceito

Antes de iniciarmos nossos experimentos com as bases de manuscritos, utilizaremos uma base de dados sintética, com o intuito de fazer uma prova de conceito. As características visualizadas nesta base são geradas através dos parâmetros média $\mu$ e variância $\sigma$ de uma distribuição normal e, correspondem a coordenadas $(X, Y)$ do plano cartesiano. Utilizaremos duas características, pois podemos visualizar melhor a movimentação dos dados neste espaço. Desta forma, foram gerados dados para conjuntos de treinamento e de testes, nos quais cada conjunto possui dez classes e cada classe é composta por quinze instâncias. A Figura 5.8 representa as características empregadas no conjunto de treinamento e no conjunto de teste. Neste trabalho, as classes representam os escritores, enquanto as instâncias representam os descritores de textura empregados.

Como podemos perceber, ambas as figuras apresentam classes muito bem comportadas

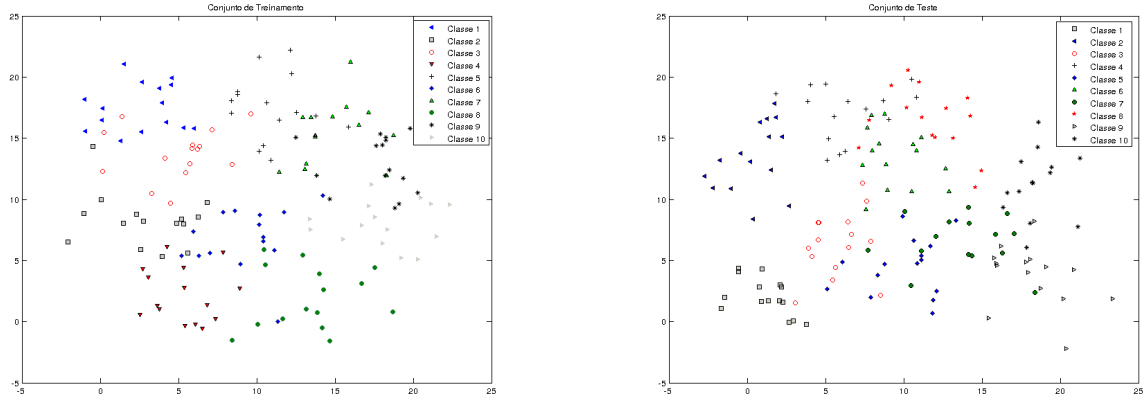



(a) (b)

Figura 5.8: Amostras da base sintética sem sobreposição. A Figura (a) representa o conjunto de treinamento, enquanto a Figura (b) representa o conjunto de teste.

e totalmente separáveis, linearmente. Na prática, dificilmente temos cenários parecidos, na maioria das vezes existe sobreposição entre as instâncias de diferentes classes, tornando árduo o processo de classificação. Todavia, para elaboração desta tese, necessitamos avaliar e compreender a relação entre as classes selecionadas. Nosso principal interesse em cenários com distribuição normal com classes totalmente separáveis é verificar se existem classes mais indicadas a fazer parte de um novo conjunto ou devido à distribuição empregada não há melhores classes. Assim, diferentes conjuntos de dados sintéticos foram avaliados, variando os parâmetros média $\mu$ e variância $\sigma$ da distribuição normal. Desta forma, geramos dados sintéticos semelhantes aos reais, de maneira que existissem sobreposição entre as classes. Assim, podemos observar a abordagem de seleção nos dois casos, naquele em que há sobreposição de dados, e em casos no qual não temos sobreposição.

Todas as etapas descritas na Figura 4.1 (*vide* Seção 4) são necessárias para o funcionamento da aplicação. Utilizamos a abordagem de dissimilaridade com cinco referências, tanto para o conjunto de treinamento quanto para o conjunto de teste ($R = S = 5$). Nestes experimentos, empregaremos o classificador SVM, juntamente com a regra da soma, utilizada para combinar as saídas do classificador. Como método de busca, utilizaremos Algoritmos Genéticos para selecionar classes (as quais representam escritores), para compor o conjunto de treinamento. Utilizaremos a Taxa de Acerto Global como função objetivo. Nestes experimentos, conforme ilustram as Figuras 5.8 e 5.9 temos 10 classes no conjunto de treinamento e 10 classes no conjunto de testes. Algoritmos genéticos são baseados na representação de bit, com cruzamento de um ponto, mutação *bit-flit* e método de amostragem estocástica uniforme (com elitismo). Os seguintes parâmetros foram empregados: população = 20, número de gerações = 100, probabilidade de cruzamento = 0,8 e probabilidade de mutação = 0.01. Os parâmetros utilizados no AG foram definidos

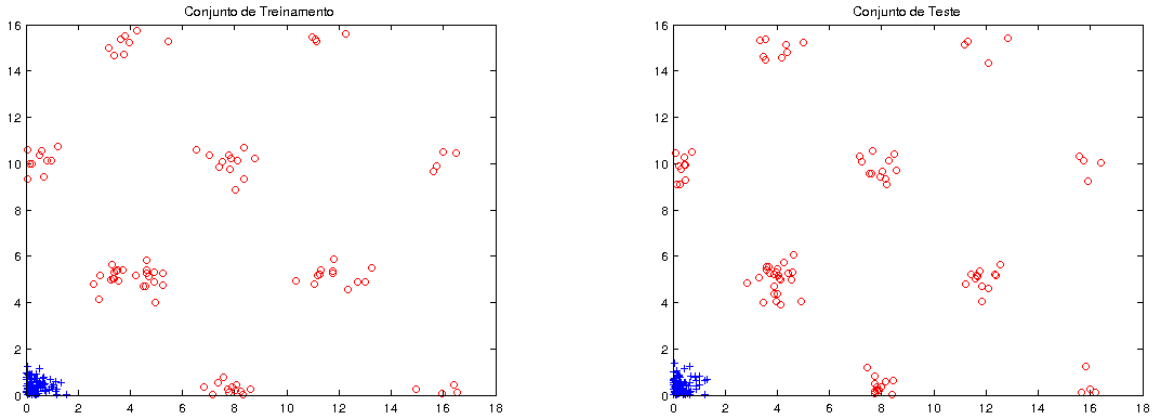



(a) (b)

Figura 5.9: Amostras da base sintética com sobreposição: a Figura (a) representa o conjunto de treinamento, enquanto a Figura (b) representa o conjunto de teste.

empiricamente, de acordo com as análises realizadas durantes os experimentos.

Através das características apresentadas na Figura 5.8, representamos as mesmas em um espaço de dissimilaridade, Figura 5.10. As Figuras 5.10(a) e 5.10(b) apresentam os conjuntos de treinamento e testes no espaço de dissimilaridade. Podemos perceber que, devido às classes possuírem boa separabilidade no espaço de características, no espaço de dissimilaridade temos, também, uma boa separação entre amostras intraclasses e interclasses.

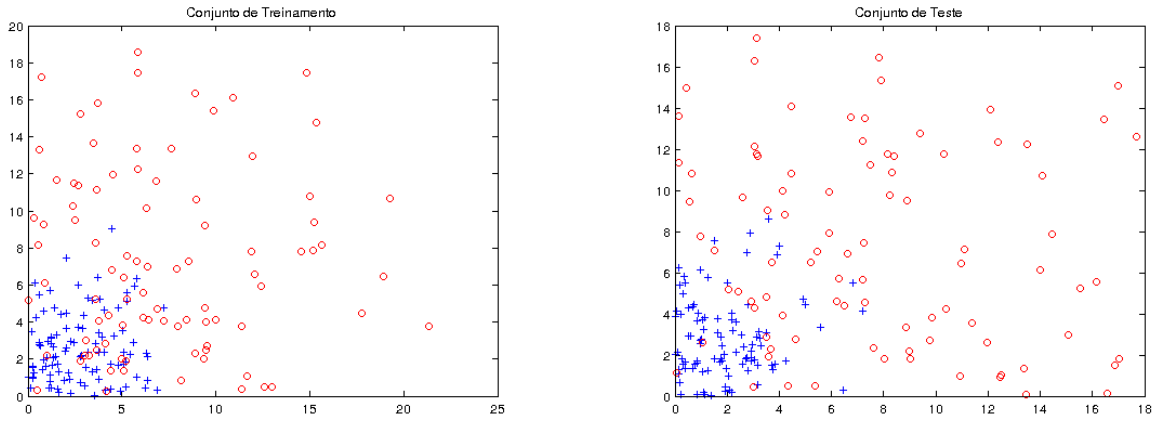



(a) (b)

Figura 5.10: Amostras da base sintética sem sobreposição no espaço de dissimilaridade: a Figura (a) representa o conjunto de treinamento, enquanto a Figura (b) representa o conjunto de teste.

Na Figura 5.11 temos a representação do espaço de dissimilaridade dos casos em

que existe sobreposição de classes e instâncias. Esta figura apresenta a transposição da Figura 5.9 para um espaço de dissimilaridade. Podemos perceber que as duas classes não são separáveis facilmente. Desta forma, daremos início aos experimentos sem e com sobreposição.




(a) (b)

Figura 5.11: Amostras da base sintética com sobreposição no espaço de dissimilaridade: a Figura (a) representa o conjunto de treinamento, enquanto a Figura (b) representa o conjunto de teste.

#### 5.3.1.1 Sem Sobreposição entre Classes

Nos experimentos realizados com a base sintética sem sobreposição, alcançamos taxas de acerto no processo de identificação de 100%, utilizando todas as classes para o treinamento. Utilizando a abordagem de seleção de escritores alcançamos as mesmas taxas, porém, utilizando um número reduzido de classes.

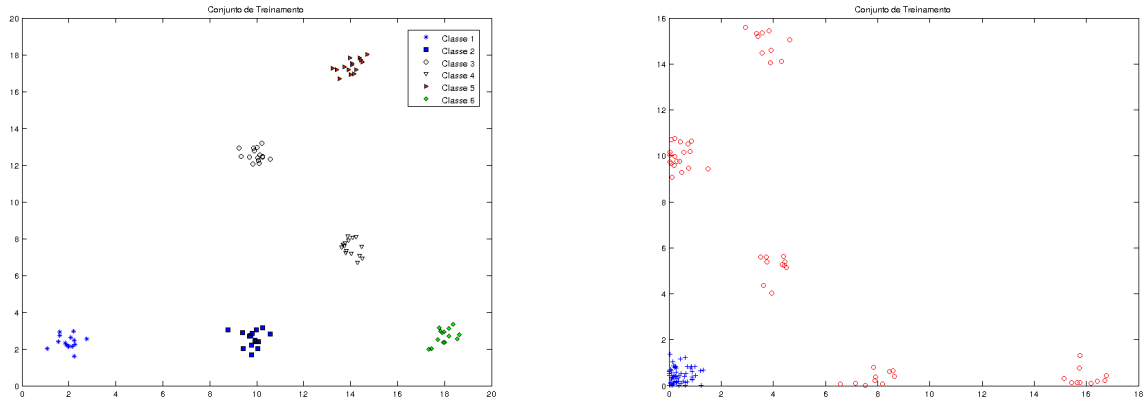
Devido aos AG's serem algoritmos de busca estocástico, realizamos dez vezes o mesmo experimento. Através destes experimentos iniciais foi possível observar que o algoritmo converge rapidamente para um ótimo global. Em todos os experimentos alcançamos taxas de acerto de 100%, entretanto, as classes selecionadas e o número de classes variaram bastante (Tabela 5.14). Percebemos que algumas classes foram selecionadas mais vezes; outras, menos vezes, mas todas as classes foram selecionadas, pelo menos, uma vez. Isso ocorre devido a todas as classes possuírem a mesma distribuição, desta forma, não existem classes melhores ou piores, todas são classes candidatas. Aparentemente, não existe uma relação de que são selecionadas as classes que se encontram em pontos opostos do espaço cartesiano ou classes muito próximas. O número de classes selecionadas também variou bastante, sendo, no mínimo, três classes selecionadas e, no máximo, sete. Desta forma, concluímos que, possuindo um espaço de características muito bem comportado, sem sobreposição, um subconjunto de classes selecionadas aleatoriamente é suficiente para gerar

um modelo robusto. A Tabela 5.14 apresenta as classes selecionadas em cada execução do experimento.

Tabela 5.14: Classes selecionadas em cada repetição.

| Experimento | Classes |
|---|---|
| 1 | 1 2 3 7 8 |
| 2 | 1 2 4 6 7 |
| 3 | 2 3 4 5 8 9 10 |
| 4 | 1 3 4 6 7 9 |
| 5 | 3 4 5 6 7 8 |
| 6 | 3 4 5 6 8 10 |
| 7 | 5 6 8 |
| 8 | 4 9 10 |
| 9 | 1 7 8 |
| 10 | 1 4 7 |

A Figura 5.12 demonstra as classes selecionadas através da abordagem de seleção, no espaço de características e de dissimilaridade.




(a) (b)

Figura 5.12: Em (a), apresentamos as classes selecionadas através do espaço de características; em (b), temos a transposição para o espaço de dissimilaridade.

Através destes experimentos conseguimos observar que para bases nas quais não há sobreposição de dados, a seleção de classes para compor o conjunto de treinamento é interessante, pois se consegue reduzir consideravelmente o número de classes no processo de treinamento, alcançando os mesmos resultados. Nosso objetivo agora é simular um comportamento próximo ao real, assim, avaliaremos o impacto da abordagem utilizando uma base sintética, na qual exista sobreposição entre as classes.

#### 5.3.1.2 Com Sobreposição entre as Classes

Experimentos considerando a sobreposição de dados entre classes foram realizados de maneira similar aos apresentados na seção anterior. Inicialmente, avaliamos o desempenho do método através da base de teste, utilizando o modelo treinado com as 10 classes (Figura 5.11). Neste experimento, a taxa de acerto alcançada foi de 90%, um desempenho de 10 pontos percentuais abaixo do experimento anterior. Isso se deve à proximidade entre as classes e à dispersão dos dados, gerando uma sobreposição entre as classes. Utilizando a abordagem de seleção de escritores, percebemos uma rápida convergência do algoritmo e uma grande redução no número de classes selecionadas, se comparado ao experimento anterior. Através da abordagem de seleção de escritores, conseguimos manter as taxas de acerto em 90%, reduzindo o número no subconjunto de treinamento. A Tabela 5.15 apresenta os resultados das classes selecionadas nas 10 repetições do experimentos.

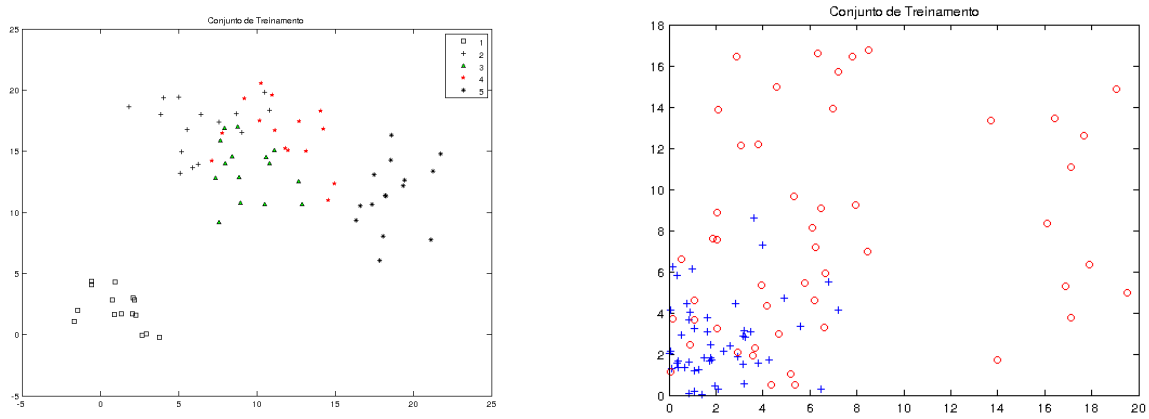
Tabela 5.15: Classes selecionadas em cada repetição.

| Experimento | Classes |
|---|---|
| 1 | 1 4 |
| 2 | 5 8 |
| 3 | 5 6 |
| 4 | 1 10 |
| 5 | 4 9 |
| 6 | 4 5 |
| 7 | 1 8 |
| 8 | 7 8 |
| 9 | 1 2 5 8 |
| 10 | 1 4 8 |

Neste e em outros experimentos realizados, variando os parâmetros média $\mu$ e variância $\sigma$ da distribuição normal, com intuito de avaliarmos o impacto da sobreposição existente entre as classes, percebemos que em cenários em que existe sobreposição entre os dados, há uma redução no número de classes eleitas para compor o conjunto de treinamento. Podemos notar, através destes resultados, que o número de classes selecionadas foi bem menor, isto se deve à sobreposição entre as classes, pois, devido à sobreposição entre os dados há uma diminuição do número de classes potenciais para compor o conjunto de treinamento. Observando as classes selecionadas através do espaço de características na Figura 5.13(a), identificar uma relação entre essas classes eleitas não é uma tarefa fácil. Observando o espaço de dissimilaridade na Figura 5.13(b), gerado através das classes selecionadas , percebemos que existe uma melhor separabilidade nas amostras intraclasses e interclasses em comparação ao uso de todas as classes no conjunto de treinamento, visto na Figura 5.11(a).

Podemos concluir que as classes foram selecionadas, pois maximizam a taxa de acerto. Assim, através da abordagem de seleção de escritores foi possível gerar um subconjunto

com menos confusão no espaço de dissimilaridade. Supõe-se, então, que uma boa separabilidade, neste espaço, é essencial para a construção de um modelo robusto. A Figura 5.13 reforça a hipótese levantada.




(a) (b)

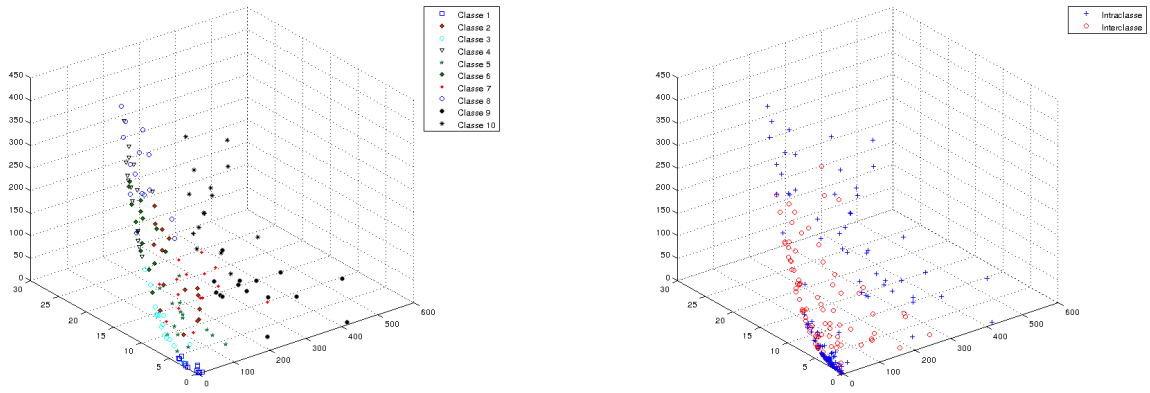
Figura 5.13: A Figura (a) apresenta as classes selecionadas através do espaço de características. Na Figura (b) temos a transposição para o espaço de dissimilaridade.

Estes experimentos demonstraram que, utilizando uma base sintética, a abordagem de seleção de escritores com intuito de gerar um modelo robusto é bastante promissora. Entretanto, há um detalhe no classificador SVM que deve ser levado em consideração. Na abordagem utilizada pelo SVM, o espaço de características original é mapeado em um espaço de características de mais alta dimensão ($x \mapsto \Phi(x)$), para que as classes possam ser linearmente separáveis. Intuitivamente, isto é "distorcer" o espaço geométrico ou inserir novas dimensões. A fim de observarmos o que ocorre em uma dimensão maior, utilizamos a função apresentada na Equação 5.2 para converter um espaço de duas dimensões para três dimensões ($\Re^2 \rightarrow \Re^3$), similar ao método *kernel trick* empregado pelo SVM.

$$(x_1, x_2) \longmapsto (z_1, z_2, z_3) = (x_1^2, \sqrt{2x_1x_2}, x_2^2) \tag{5.2}$$

Assim, para melhor compreensão das classes selecionadas pelo SVM, mudamos a apresentação no espaço de características de duas para três dimensões. Neste experimento, consideramos somente a base sintética com sobreposição. A Figura 5.14 refere-se às mesmas características anteriormente demonstradas, através de um espaço 2D apresentado na Figura 5.9, porém, neste caso, representada através de três dimensões. Podemos perceber que as mesmas não são separáveis linearmente em uma dimensão mais alta. Isto também ocorre para o espaço de dissimilaridade, podendo ser observado através da Figura 5.14(b).

Através destas demonstrações, podemos concluir que, mesmo o SVM não utilizando o espaço original das características, e sim, um espaço mapeado, podemos notar que em tal

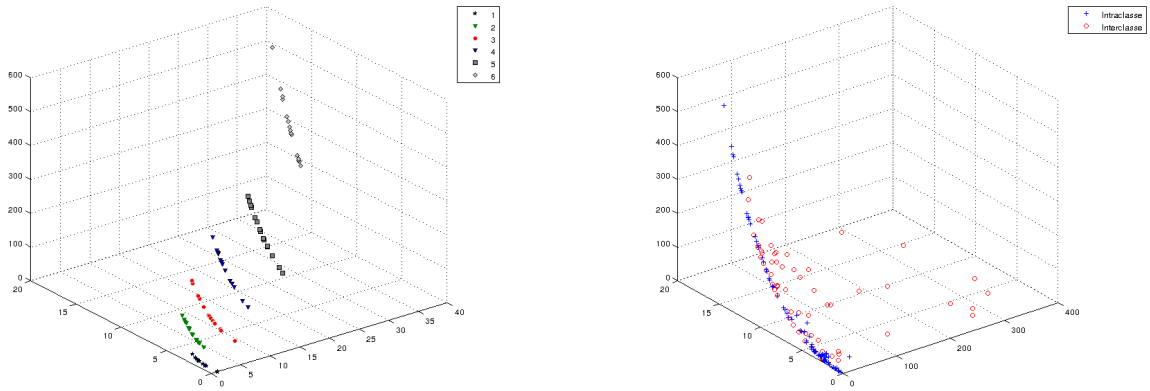



(a) (b)

Figura 5.14: A Figura (a) apresenta as 10 classes através do espaço de características. Na Figura (b) temos a transposição para o espaço de dissimilaridade.

dimensão, estas características não são separáveis. Percebemos ainda que, no espaço de dissimilaridade também há certo nível de confusão entre as classes. Notamos ainda que elas não são totalmente separáveis, mesmo em uma dimensão maior.

Através das classes selecionadas, representadas em um espaço tridimensional na Figura 5.15, podemos notar algumas características interessantes, por exemplo, através da seleção temos classes totalmente separáveis em uma dimensão superior. Isto colabora para gerar um espaço de dissimilaridade, no qual as amostras intraclasses fiquem mais concentradas, evitando uma maior confusão.




(a) (b)

Figura 5.15: A Figura (a) apresenta as classes selecionadas através do espaço de características em 3D. Na Figura (b) temos a transposição para o espaço de dissimilaridade representado em três dimensões.

#### 5.3.1.3 Análise dos Experimentos

Em suma, através dos experimentos utilizando a base sintética, conseguimos perceber algumas características, como:

- A proposta de seleção de escritores para compor o conjunto de treinamento apresentou-se promissora para uma base sintética;
- Aparentemente em um ambiente com boa separabilidade entre as classes, o processo de seleção de escritores não tem um impacto forte, pois classes eleitas aleatoriamente podem gerar um modelo robusto, entretanto, podemos reduzir consideravelmente a quantidade de escritores presentes do conjunto de treinamento;
- Utilizando um conjunto, no qual existe sobreposição entre as classes, podemos notar que houve uma melhora no desempenho do sistema, reduzindo consideravelmente o número de classes no conjunto de treinamento;
- As classes selecionadas têm forte relação ao espaço de dissimilaridade. Aparentemente não é necessária uma ótima margem de separação entre as duas classes no espaço de dissimilaridade para gerar um bom modelo;
- Em suma, muitas das classes presentes em um conjunto de treinamento não colaboram para a criação de um modelo robusto, assim, reduzindo o número de classes podemos gerar modelos mais robustos, ou então, reduzir consideravelmente o número de classes no conjunto de treinamento, consequentemente, reduzindo o custo para criar um modelo.

### 5.3.2 Experimentos Utilizando Bases de Manuscritos

Apresentaremos, a seguir, os experimentos realizados com seleção de escritores com objetivo de verificar o desempenho e a redução do número de escritores após o uso desta abordagem. Através de um esquema de classificação, embasado na abordagem escritor-independente usando a dissimilaridade, propomos um método que visa reduzir o número de escritores presentes no conjunto de treinamento para geração de um modelo. Buscamos, ainda, alcançar possíveis melhorias nas taxas de acertos, pois através dos experimentos anteriores, percebemos que não há necessidade de um grande número de escritores para gerar um modelo eficiente, mas sim, de escritores que, combinados, gerem um modelo robusto.

Em todos os nossos experimentos utilizamos os descritores de textura $LBP^{U2}_{8,2}$ e LPQ com janela $7 \times 7$. Fixamos o número de escritores nos conjuntos de testes, seguindo trabalhos anteriores, apresentados por Hanusiak et al. [42] e Bertolini et al. [12]. Assim, 115, 240 e 90 escritores foram utilizados para as bases BFL, IAM e *Firemaker*, respectivamente.

A partir das partições geradas nos experimentos descritos na seção 5.1, utilizamos o menor conjunto de treinamento proposto para as bases BFL, IAM e *Firemaker*. Nos experimentos empregando o maior conjunto de treinamento, alcançamos as mesmas taxas de acerto. Outro fato interessante foi que, utilizando um grande conjunto de treinamento, havia uma redução de cerca de 50% do número de escritores, porém, como havia muitos escritores, o conjunto ainda permanecia grande. Em decorrência da quantidade de escritores, tínhamos ainda um alto custo computacional.

A abordagem de seleção de escritores é apresentada na Figura 4.13 (*vide* seção 4.6). Assim, foram utilizados dois subconjuntos de validação, denominados *Validação 1* e *Validação 2*. O subconjunto *Validação 1* foi utilizado para calcular a função de aptidão do Algoritmo Genético. O subconjunto *Validação 2* foi gerado com o objetivo de evitar *over-fitting*. Por fim, através dos escritores selecionados, verificamos as taxas de acerto global em função do conjunto de *Teste*. No esquema de dissimilaridade, usamos o mesmo número de referências para os conjuntos de treinamento e teste $R = S = [3, 5, 9]$. Em todos os nossos experimentos, o classificador *Support Vector Machine* e a regra da Soma foram utilizados. Utilizamos a regra da Soma, pois a regra da Mediana apresentou taxas mais baixas, isto pode ser notado nos experimentos anteriores, utilizando poucos escritores. A taxa de acerto global foi utilizada como função objetivo, entretanto, avaliamos a AUC como função objetiva, a qual não apresentou melhores resultados. Os parâmetros empregados nos Algoritmos Genéticos foram os mesmos apresentados na seção anterior (*vide* seção 5.3.1). As taxas apresentadas a seguir referem-se à média aritmética de três repetições juntamente com seus respectivos desvios padrões.

Inicialmente, avaliaremos o ponto de convergência do Algoritmo genético. Tal experimento visa observar alguns aspectos da abordagem de seleção. Em seguida, realizaremos experimentos com as bases BFL, IAM e *Firemaker*. A fim de avaliar a robustez da abordagem de seleção de escritores, utilizaremos as três bases em conjunto para avaliar o quão robusto é nosso sistema para diferentes línguas.

Utilizamos para estes experimentos a base BFL com $R = S = 5$. Temos como propósito, avaliar o número de gerações necessárias para que haja convergência para um ótimo global e ainda verificar se, ao melhorar as taxas no subconjunto de *Validação 1*, temos melhoria de desempenho nos subconjuntos de *Validação 2* e *Teste*. Através de experimentos realizados, podemos notar que, com poucas gerações, o algoritmo converge para um bom resultado. Outra característica percebida é que, com muitas gerações, o modelo gerado passa a ser específico para o conjunto de *Validação 1*, o que o torna muito especialista para aquele subconjunto, não havendo melhoras nas taxas dos outros subconjuntos. Para criar um conjunto que seja generalista, necessitamos de poucas gerações no AG. Os resultados alcançados, utilizando a abordagem proposta, referem-se ao melhor ponto operacional apresentado pelo subconjunto de *Validação 2*.

Experimentos visando otimizar diretamente o conjunto de *Teste*, através de AG, demonstraram um ganho maior de desempenho. Entretanto, otimizando o conjunto de Teste,

estamos criando um classificador especialista para um determinado conjunto, sendo que, nesta abordagem, desejamos um classificador genérico.

#### 5.3.2.1 Experimentos Usando a Base BFL

Experimentos, utilizando a abordagem de escritor-independente com 25 escritores no conjunto de treinamento, reportam taxas de acertos de 89,65%, ($R = S = 3$), usando descritor LBP. Através da abordagem proposta, alcançamos taxa de 95,1% de identificação usando 9,6 dos 25 escritores empregados na abordagem anterior, através dos mesmos parâmetros. Podemos observar que houve ganho de 5 pontos percentuais nas taxas de acerto global, demonstrando, neste caso, um ganho de desempenho considerável. Percebemos, também, uma redução de 56% do número de escritores.

A Tabela 5.16 nos mostra que em todos os casos houve melhoras nas taxas de acerto. Podemos perceber que a abordagem proposta tem um maior impacto ao utilizarmos um número pequeno de referências ($R = S = 3$ ou 5). É provável que tal comportamento aconteça devido ao espaço de dissimilaridade. Utilizando menos referências, é possível que o modelo gerado seja mais genérico, tornando-o mais robusto para o processo de identificação. Podemos observar que, ao aumentar o número de referências, temos uma melhora nas taxas de acerto, aumentando em quase 4 pontos percentuais as taxas de acerto. Notamos também que houve uma redução entre 32 e 56% do número de escritores. Em geral, o descritor LPQ apresentou resultados melhores que o LBP. Lembrando que, para todos os resultados apresentados, foi considerada a média de três repetições. Na abordagem sem seleção, foram utilizados 25 escritores no conjunto de treinamento.

| Descritor | Ref. | Com Seleção | | | | Sem Seleção |
|---|---|---|---|---|---|---|
| | R = S | Taxa (%) | $\sigma$ | Escritores | $\sigma$ | Taxa (%) |
| LBP | 3 | 95,1 | 0,005 | 9,6 | 2,3 | 89,5 |
| | 5 | 95,7 | 0,015 | 12 | 1,0 | 94,7 |
| | 9 | 98,0 | 0,010 | 14 | 4,3 | 99,8 |
| LPQ | 3 | 95,5 | 0,010 | 11,3 | 0,57 | 96,5 |
| | 5 | 98,3 | 0,009 | 13 | 3,46 | 99,1 |
| | 9 | 99,4 | 0,010 | 14 | 1,0 | 99,9 |

Tabela 5.16: Taxas de Acerto Global (%) utilizando a abordagem de seleção de escritores - base BFL.

Fica claro na Tabela 5.16 que, utilizando o descritor LBP combinado com a abordagem de seleção de escritores, conseguimos uma melhora nos resultados. Isto não ocorre nos experimentos com LPQ. O motivo disso pode ser o bom desempenho do descritor de textura LPQ, que gera modelos mais robustos, independente de selecionar ou não escritores. A Figura 5.16 apresenta os escritores selecionados em um caso específico (escritores selecionados em cinza).

Com o intuito de melhorarmos as taxas de acerto global, a população inicial contou


Figura 5.16: Escritores selecionados utilizando a base BFL, descritor LPQ e $R = S = 5$.

com escritores selecionados em experimentos anteriores. Neste caso, mesmo inicializando o algoritmo com uma solução ótima, o AG não convergiu para resultados melhores. Assim, os escritores selecionados continuaram os mesmos.

Como descrito anteriormente, foram realizados experimentos utilizando um conjunto de treinamento com todos os escritores que não fizeram parte dos conjuntos de *Teste*, *Validação 1* e *Validação 2*, ou seja, 80 escritores. Verificamos que, neste caso, não houve nenhuma melhoria nas taxas.

#### 5.3.2.2 Experimentos Usando a Base IAM

Em nossos experimentos, usando a base IAM, empregamos o mesmo protocolo usado nos experimentos anteriores. Todavia, devido à diferença de tamanho da base, utilizamos uma quantidade maior de escritores nos subconjuntos. Assim, temos os seguintes números de escritores para cada subconjunto: 50 no Treinamento, 125 na *Validação 1*, 125 na *Validação 2* e 240 escritores no conjunto de *Teste*.

Observando a Tabela 5.17, percebemos que as taxas apresentadas estão bem abaixo das apresentadas pela base BFL. Várias características podem ter influenciado, como o fato desta base ser texto-independente, resultando em uma quantidade de escrita bastante variável, o que gera menos textura para representar um escritor. Outro fator, é o número de escritores no conjunto de teste, como temos um número maior de escritores no teste, fica difícil tal comparação. Contudo, nosso objetivo é verificar o impacto da abordagem proposta na base IAM, sendo esta uma base desafiadora e de texto-independente.

Assim, na Tabela 5.17, é fácil notar que temos uma melhoria de desempenho considerável ao aumentarmos o número de amostras de referência, $R$ e $S$. Em ambos os descritores, houve uma melhoria considerável; 21 pontos percentuais usando o LBP e 15 pontos percentuais empregando o LPQ. Podemos notar a superioridade do LPQ quando utilizamos menos referências. Neste caso, houve uma redução de até 56% no número de escritores.

Avaliando os experimentos com a base IAM, é possível notar uma melhoria em todos os casos ao empregar a abordagem de seleção de escritores. Diferente da base BFL, a base IAM apresentou um melhor desempenho ao selecionar escritores para compor o conjunto de treinamento. Contudo, em ambas as bases, vimos uma redução no número de escritores para compor o conjunto, demonstrando que não necessitamos de todos os escritores para gerar um modelo robusto. Aparentemente, ao utilizar um número menor de referências para $R$ e $S$ criamos um modelo mais genérico, contribuindo para um maior ganho de desempenho quando comparados as abordagens com e sem seleção de escritores.

| Descritor | Ref. | Com Seleção | | | | Sem Seleção |
|---|---|---|---|---|---|---|
| | R = S | Taxa (%) | $\sigma$ | Escritores | $\sigma$ | Taxa (%) |
| LBP | 3 | 68,2 | 0.035 | 26.0 | 2.64 | 60,0 |
| | 5 | 76,5 | 0.025 | 25.6 | 8.38 | 75,0 |
| | 9 | 91,3 | 0.013 | 28.6 | 1.15 | 91,0 |
| LPQ | 3 | 77,5 | 0.081 | 26.3. | 1.15 | 75,0 |
| | 5 | 81,8 | 0.024 | 22.0 | 1.73 | 77,0 |
| | 9 | 93,1 | 0.012 | 27.3 | 1.15 | 92,0 |

Tabela 5.17: Taxas de Acerto Global (%) utilizando a abordagem de seleção de escritores - base IAM.

#### 5.3.2.3 Experimentos Usando a Base *Firemaker*

Como descrito anteriormente, a base *Firemaker* pode ser trabalhada de diversas maneiras, já que a mesma consta de quatro cartas por escritor, sendo: texto-dependente, texto-independente, caixa alta e falsificação. Neste experimento, utilizamos uma única carta com texto-dependente. As taxas de acertos apresentadas são próximas às apresentadas utilizando a base BFL. Utilizamos os seguintes números de escritores para cada subconjunto: *Validação 1* = 45, *Validação 2* = 45, *Teste* = 90 sendo o conjunto de treinamento composto por 20 escritores. Resultados dos experimentos usando a base *Firemaker* são descritos na Tabela 5.18. Observando os resultados apresentados pela base BFL (Tabela 5.16) percebe-se que são próximos aos apresentados pela *Firemaker*. Desta forma, é possível notar que texto-dependente tende a apresentar melhores taxas.

| Descritor | Ref. | Com Seleção | | | | Sem Seleção |
|---|---|---|---|---|---|---|
| | R = S | Taxa (%) | $\sigma$ | Escritores | $\sigma$ | Taxa (%) |
| LBP | 3 | 96,7 | 0.019 | 10.3 | 1.5 | 94,4 |
| | 5 | 91,9 | 0.006 | 9.3 | 1.5 | 91,1 |
| | 9 | 96,7 | 0.000 | 12.3 | 0.5 | 97,7 |
| LPQ | 3 | 98,1 | 0.006 | 8.6 | 0.57 | 96,6 |
| | 5 | 98,9 | 0.011 | 11.6 | 4.9 | 96,6 |
| | 9 | 97,8 | 0.019 | 10.3 | 1.5 | 98,8 |

Tabela 5.18: Taxas de Acerto Global (%) utilizando a abordagem de seleção de escritores - base *Firemaker*.

Um segundo experimento foi realizado com intuito de verificar o impacto ao possuirmos texto-dependente, porém, com diferentes línguas escritas, neste caso, o Português e Holandês. Para isso, unimos as bases BFL e *Firemaker*. Usamos percentuais de escritores para os subconjuntos de *Validação 1*, *Validação 2* e Teste similar aos empregados em experimentos com outras bases, ficando com 105, 105 e 205, respectivamente e, no conjunto de treinamento, foram utilizados 45 escritores. Através destes experimentos, poderemos avaliar o desempenho do sistema possuindo somente texto-dependente e uma maior quantidade de escritores no teste.

Através da Tabela 5.19 percebemos dois fatores interessantes: primeiro, a abordagem

| Descritor | Ref. | Com Seleção | | | | Sem Seleção |
|---|---|---|---|---|---|---|
| | R = S | Taxa (%) | $\sigma$ | Escritores | $\sigma$ | Taxa (%) |
| LBP | 3 | 86,3 | 0.034 | 19.6 | 3.2 | 77,0 |
| | 5 | 84,6 | 0.023 | 20.0 | 4.0 | 83,9 |
| | 9 | 94,6 | 0.020 | 24.3 | 3.2 | 93,1 |
| LPQ | 3 | 89,3 | 0.039 | 16.6 | 1.5 | 81,9 |
| | 5 | 87,0 | 0.003 | 20.6 | 3.7 | 86,8 |
| | 9 | 94,3 | 0.010 | 22.3 | 2.0 | 95,1 |

Tabela 5.19: Taxas de Acerto Global (%) utilizando a abordagem de seleção de escritores - base BFL + *Firemaker*.

proposta utilizando textura é robusta ao possuirmos escrita em diferentes línguas. Segundo, mesmo neste caso, no qual havia duas línguas, a utilização de texto-dependente apresenta melhor desempenho que a texto-independente, principalmente, ao empregarmos um número pequeno de referências ($R = S = 3$). A redução do número de escritores mantém-se semelhante aos experimentos anteriores. Lembrando que a base BFL possui amostras de 315 escritores, enquanto a *Firemaker* contém 250 escritores. Desta forma, cada subconjunto possui um número de escritores proporcional ao tamanho da sua base.

#### 5.3.2.4 Experimentos Usando a Base BFL + IAM + *Firemaker*

Por fim, unimos as três bases de dados utilizadas neste trabalho, BFL, IAM e *Firemaker* a fim de verificar o impacto ao possuirmos uma base mista, com texto-dependente e texto-independente, e uma grande quantidade de escritores. Neste caso, utilizamos blocos de textura de tamanho 256 × 128. Neste experimento, temos 650 escritores com texto-independente e 565 com texto-dependente. Foram utilizados 95 escritores para o conjunto de treinamento, 445 escritores no conjunto de *Teste* e 205 escritores para os conjuntos de *Validação 1* e *Validação 2*. A Tabela 5.20 apresenta as taxas de acerto empregando a abordagem proposta.

| Descritor | Ref. | Com Seleção | | | | Sem Seleção |
|---|---|---|---|---|---|---|
| | R = S | Taxa (%) | $\sigma$ | Escritores | $\sigma$ | Taxa (%) |
| LBP | 3 | 71,1 | 0.017 | 49.6 | 7.5 | 67,4 |
| | 5 | 81,1 | 0.027 | 50 | 3.6 | 81,5 |
| | 9 | 91,8 | 0.008 | 50 | 1.4 | 92,3 |
| LPQ | 3 | 76,6 | 0.015 | 49 | 3.6 | 68,0 |
| | 5 | 81,1 | 0.025 | 45.6 | 5.0 | 84,0 |
| | 9 | 92,1 | 0.009 | 48.5 | 0.7 | 93,2 |

Tabela 5.20: Taxas de Acerto Global (%) utilizando a abordagem de seleção de escritores - base BFL + IAM + *Firemaker*.

Observando a Tabela 5.20, fica claro a melhora no desempenho ao aumentar o número de referências $R$ e $S$. Outra característica é a redução de cerca de 50% no número de escritores para compor o conjunto de treinamento. Podemos notar que, usando poucas

referências, $R = S = 3$, o ganho de desempenho utilizando a abordagem proposta é considerável. Entretanto, ao aumentarmos as referências, temos como vantagem a redução do número de escritores para compor o conjunto de treinamento, porém, através do modelo gerado, temos taxas de 0,5 a 1,1 pontos percentuais piores do que utilizando todos os escritores no conjunto de treinamento.

Mesmo assim, esta abordagem demonstra boas taxas de acertos ao utilizarmos bases com manuscritos em diferentes línguas. Percebemos, também, a robustez desta, utilizando um grande conjunto de escritores no teste.

## 5.4 Considerações sobre os Experimentos

O trabalho desenvolvido, até o momento, abordando verificação e identificação de escritor, utiliza abordagem similar à proposta por Hanusiak et al. [42], de forma que sua principal vantagem é evitar a segmentação, pois transforma o texto manuscrito em textura para representar o escritor. Nesta abordagem, empregamos descritores de textura para extrair características, de maneira que o vetor de características extraído é transformado em um vetor de dissimilaridades, o qual é utilizado como entrada no classificador SVM.

Experimentos com LBP e LPQ demonstraram que o LPQ é um descritor de textura robusto para esta aplicação. Percebemos, também, que o mesmo supera, com vantagem, o GLCM no processo de identificação. No processo de verificação, percebemos um menor impacto em relação aos descritores.

A partir de nossos resultados, empregando a abordagem escritor-independente utilizando conjuntos de treinamento com números fixos de escritores, notamos que a abordagem de dissimilaridade apresentou-se como estratégia viável para problemas de verificação e identificação, já que em nossos experimentos, empregando as bases BFL e IAM alcançamos excelente desempenho, comparável ou até mesmo superior, aos descritos em literatura.

Empregando a abordagem proposta em que escritores são selecionados para compor o conjunto de treinamento com o objetivo de gerar modelos mais robustos, percebemos que a abordagem é promissora, pois através da mesma podemos notar uma redução de cerca de 50% do número de escritores para compor o conjunto de treinamento. Por fim, alcançamos taxas de acertos globais superiores em determinados casos, utilizando a abordagem de seleção de escritores, superando as taxas apresentadas através do método escritor-independente.

# CAPÍTULO 6

# CONCLUSÕES

O principal objetivo deste trabalho foi construir uma aplicação de identificação e verificação de escritor que possa auxiliar peritos nesta tarefa. Para isso, utilizamos diversas abordagens, visando a uma melhoria no desempenho do sistema. A combinação destas diversas técnicas contribui para a originalidade deste trabalho, demonstrando a eficiência da textura para identificação de escritor e o desempenho do uso da abordagem de escritor-independente. Contudo, a proposta inovadora deste trabalho é a seleção de escritores para compor um conjunto de treinamento, a fim de reduzir escritores neste conjunto e melhorar o desempenho do sistema. Através dos resultados demonstrados nesta tese, podemos concluir que obtivemos sucesso na tarefa proposta, pois, avaliando a proposta em diversas bases de dados, conseguimos alcançar taxas de acerto que superam as descritas em literatura.

Inicialmente, conseguimos constatar que a utilização da textura é vantajosa para o processo, pois não há necessidade de segmentação, tarefa com alto custo computacional e alta complexidade, tornando-se independente de língua. Percebemos também que, para o processo de identificação, a abordagem de geração de textura foi essencial para o desempenho do sistema. Outro fator que contribuiu para o sucesso da aplicação foram os descritores de textura empregados nesta pesquisa. Percebemos que mesmo o descritor GLCM tendo apresentado bons resultados para a verificação de escritor, para o processo de identificação, este não apresentou resultados satisfatórios. Através dos descritores LBP e LPQ, conseguimos melhorar, consideravelmente, as taxas de identificação. Isto demonstra a robustez destes descritores, principalmente, do descritor LPQ, responsável pelas melhores taxas alcançadas. Mesmo sendo um descritor desenvolvido para ser ótimo com imagens borradas, o mesmo apresentou excelentes resultados utilizando imagens sem ruídos. O LPQ apresentou ótimo desempenho utilizando um número baixo de referências ($R = S = 3$). O número de referência $R$ e $S$ também foi decisivo para melhorarmos o desempenho do sistema, pois percebemos que o número de referências tem um forte impacto nas taxas de identificação, impacto maior que o número de escritores. Quanto ao número de escritores notamos que:

- O número de escritores no conjunto de treinamento tem um impacto maior no processo de identificação de escritor. Na verificação, este impacto é menor;
- A lógica, quanto mais escritores no conjunto de treinamento melhor, não é verdade para todos os casos;
- Com poucos escritores no conjunto de treinamento, conseguimos gerar modelos ro-

bustos;

- Através da abordagem de seleção de escritores, conseguimos melhorar as taxas de acerto global principalmente ao possuir poucas referências;
- Com um número reduzido de escritores no conjunto de treinamento é possível atingir taxas muito próximas às taxas utilizando todo um conjunto, cerca de 8 a 10 vezes o número inicial;

Com relação às bases de dados empregadas, podemos observar que a proposta apresentou-se eficiente para as três. A utilização de bases com texto-dependente apresentou desempenho próximos a quatro pontos percentuais melhores quando comparados com a base texto-independente. Em virtude do uso da textura, a proposta apresentou-se robusta para diferentes línguas escritas (Português do Brasil, Inglês e Holandês) e diferentes estilos de escrita. Por fim, analisando as taxas de acerto para Top-5 e Top-10, conseguimos resultados próximos a 100%.

## 6.1 Contribuições

Como contribuições, podemos destacar:

- Proposta de uma abordagem inovadora para seleção de escritor para compor conjuntos de treinamento, gerando modelos robustos através de um número reduzido de escritores;
- Avaliação de uma abordagem não dependente de segmentação e da língua escrita;
- Verificação do impacto de descritores de textura;
- Avaliação do desempenho do uso de diferentes números de referências ($R$ e $S$);
- Estudo do uso de regras de fusão para combinar a saída de classificadores;
- Avaliação do impacto do número de escritores no processo de identificação de escritor;
- Estudo do processo de seleção de escritores para geração de modelos robustos;
- Geração de um modelo robusto, através de poucos escritores no conjunto de treinamento;
- Abordagem independente de escritor não havendo necessidade de retreino do modelo, em caso de inserções de novos escritores no conjunto de teste;
- Uma ferramenta de apoio a peritos forenses;
- Através de uma abordagem escritor-independente, alcançamos taxas inéditas na literatura para as bases BFL, IAM e Firemaker.

## 6.2 Trabalhos Futuros

Através da realização deste trabalho pudemos observar detalhes que não faziam parte do escopo desta pesquisa, porém, podem contribuir para melhoria do desempenho da mesma.

- Verificar o impacto de novos descritores de textura;
- Analisar o impacto de possuirmos mais referências ($R$ e $S$) em algumas bases;
- Analisar técnicas para geração de textura através de manuscritos, de forma que possam gerar uma textura ainda mais forte para representar o escritor;
- Verificar a robustez da aplicação para uma base de língua Árabe ou Chinesa, na qual a escrita possui grande diferença das apresentadas aqui;
- Verificar a proposta de seleção de escritores em outras áreas, como a verificação de assinaturas.

# REFERÊNCIAS


[1] P. Almeida, L. S. Oliveira, E. Silva, A. Britto, e A. Koerich. Parking space detection using textural descriptors. UK Manchester, editor, *IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC)*, páginas 3603–3608, 2013.

[2] A. M. M. M. Amaral, C. O. A. Freitas, e F. Bortolozzi. The graphometry applied to writer identification. *IPCV'12 - International Conference on Image Processing, Computer Vision, and Pattern Recognition*, vol.1:pp.10–16, 2012. Las Vegas, USA.

[3] A. M. M. M. Amaral, C. O. A. Freitas, e F. Bortolozzi. Multiple graphometric features for writer identification as part of forensic handwriting analysis. *IPCV'13 - International Conference on Image Processing, Computer Vision, and Pattern Recognition*, 2013.

[4] A.M.M.M Amaral, C. O. A. Freitas, e F. Bortolozzi. Feature selection for forensic handwriting identification. *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, páginas 922–926, 2013.

[5] S. P. Balbás. *Reconociemiento de Escritor Independiente de Texto Basado en Características de Textura*. Tese de Doutorado, Universidad Autonoma de Madrid, 2007.

[6] F. Baranoski, L. S. Oliveira, e E. Justino. Writer identification based on forensic science approach. October 9-12 San Jose, Costa Rica, editor, *XXXIII Latin American Conference on Informatics*, páginas 25–32, 2007.

[7] F. L. Baranoski. Verificação da autoria em documentos manuscritos usando svm. Dissertação de Mestrado, Pontifícia Universidade Católica do Paraná, 2005.

[8] H. Bay, A. Ess, T. Tuytelaars, e L. V. Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346 – 359, 2008. Similarity Matching in Computer Vision and Multimedia.

[9] A. Bensefia, T. Paquet, e L. Heutte. A writer identification and verification system. *Pattern Recognition Letters*, 26(13):2080 – 2092, 2005.

[10] D. Bertolini, L. S. Oliveira, E. Justino, e R. Sabourin. Reducing forgeries in writer-independent off-line signature verification through ensemble of classifiers. *Pattern Recognition*, 43:387–396, Janeiro de 2010.

[11] D. Bertolini, L.S. Oliveira, E. Justino, e R. Sabourin. Ensemble of classifiers for off-line signature verification. *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, páginas 283 –288, 2008.

[12] D. Bertolini, L.S. Oliveira, E. Justino, e R. Sabourin. Texture-based descriptors for writer identification and verification. *Expert Systems with Applications*, 40(6):2069 – 2080, 2013.

[13] D. Bertolini, L.S. Oliveira, E. Justino, e R. Sabourin. Assessing textural features for writer identification on different writing styles and forgeries. *International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, 24-28 August*, 2014, (Aceito para Publicação).

[14] A. F. Bisquerra. *Writer Identification by a Combination of Graphical Features in the Framework of Old Handwritten Music Scores*. Tese de Doutorado, Universitat Autònoma de Barcelona, 2009.

[15] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, e A. W. Senior. The relation between the roc curve and the cmc. *4th Workshop Automatic Identification Advanced Technologies*, páginas 15–20, 2005.

[16] A. A. Brink, M. Bulacu, e L. Schomaker. How much handwritten text is needed for text-independent writer verification and identification. *19th International Conference on Pattern Recognition (ICPR 2008)*, páginas 1–4, 2008.

[17] A.A. Brink, R.M.J. Niels, R.A. van Batenburg, C.E. van den Heuvel, e L.R.B. Schomaker. Towards robust writer verification by correcting unnatural slant. *Pattern Recognition Letters*, 32(3):449 – 457, 2011.

[18] A.A. Brink, J. Smit, M.L. Bulacu, e L.R.B. Schomaker. Writer identification using directional ink-trace width measurements. *Pattern Recognition*, 45(1):162 – 171, 2012.

[19] P. Brodatz. *Textures: A Photographic Album for Artists & Designers*. New York: Dover, 1966.

[20] M. Bulacu e L. Schomaker. Text-independent writer identification and verification using textural and allographic features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(4):701 –717, 2007.

[21] M. G. Campiteli. *A Caminhada do Turista como Ferramenta na Identificação de Padrões*. Tese de Doutorado, Universidade de São Paulo, 2007.

[22] S. Cha, , e S. N. Srihari. On measuring the distance between histograms. *Pattern Recognition*, 35(6):1355–1370, junho de 2002.

[23] S. Cha e S. Srihari. Writer identification: Statistical analysis and dichotomizer. *Advances in Pattern Recognition*, volume 1876 of *Lecture Notes in Computer Science*, páginas 123–132. Springer Berlin - Heidelberg, 2000. 10.1007-3-540-44522-6 13.

[24] R. Coll, A. Fornes, e J. Llados. Graphological analysis of handwritten text documents for human resources recruitment. *10th International Conference on Document Analysis and Recognition, 2009 (ICDAR '09)*, páginas 1081–1085, Julho de 2009.

[25] S. Cong, R. Xiao-gang, e M. Tian-Lu. Writer identification using gabor wavelet. *Proceedings of the 4th World Congress on Intelligent Control and Automation*, volume 3, páginas 2061 – 2064 vol.3, 2002.

[26] Y.M.G. Costa, L.S. Oliveira, A.L. Koerich, F. Gouyon, e J.Martins. Music genre classification using lbp textural features. *Signal Processing*, 92(11):2723 – 2737, 2012.

[27] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, 2(7):1160–1169, Julho de 1985.

[28] T. G. Dietterich. Ensemble methods in machine learning. *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, páginas 1–15, London, UK, UK, 2000. Springer-Verlag.

[29] C. Djeddi, I. Siddiqi, L. Souici-Meslati, e A. Ennaji. Text-independent writer recognition using multi-script handwritten texts. *Pattern Recognition Letters*, 34(10):1196–1202, julho de 2013.

[30] L. Du, X. You, H. Xu, Z. Gao, e Y. Tang. Wavelet domain local binary pattern features for writer identification. *20th International Conference on Pattern Recognition (ICPR)*, páginas 3691 –3694, Agosto de 2010.

[31] G. Eskander, R. Sabourin, e E. Granger. A dissimilarity-based approach for biometric fuzzy vaultsapplication to handwritten signature images. Alfredo Petrosino, Lucia Maddalena, e Pietro Pala, editors, *New Trends in Image Analysis and Processing ICIAP 2013*, volume 8158 of *Lecture Notes in Computer Science*, páginas 95–102. Springer Berlin Heidelberg, 2013.

[32] G. Eskander, R. Sabourin, e E. Granger. Hybrid writer-independentwriter-dependent offline signature verification system. *IET Biometrics*, 2:169–181(12), December de 2013.

[33] G. Eskander, R. Sabourin, e E. Granger. On the dissimilarity representation and prototype selection for signature-based bio-cryptographic systems. Edwin Hancock e Marcello Pelillo, editors, *Similarity-Based Pattern Recognition*, volume 7953 of *Lecture Notes in Computer Science*, páginas 265–280. Springer Berlin Heidelberg, 2013.

[34] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 227(8):861–874, 2006.

[35] C. Freitas, L. S. Oliveira, R. Sabourin, e F. Bortolozzi. Brazilian forensic letter database. *11th International Workshop on Frontiers on Handwriting Recognition (IWFHR-11)*, 2008.

[36] U. Garain e T. Paquet. Off-line multi-script writer identification using ar coefficients. *10th International Conference on Document Analysis and Recognition, ICDAR '09.*, páginas 991 –995, Julho de 2009.

[37] S. Garcia, J. Derrac, J. R. Cano, e F. Herrera. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):417–435, 2012.

[38] R. C. Gonzalez e R. E. Woods. *Digital Image Processing*. Pearson Prentice Hall, 2008.

[39] E. Grosicki, M. Carre, J. Brodin, e E. Geoffrois. Results of the rimes evaluation campaign for handwritten mail processing. *ICDAR '09: Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, páginas 941–945, Washington, DC, USA, 2009. IEEE Computer Society.

[40] Z. Guo, L. Zhang, e D. Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions Image Processing*, páginas 1657–1663, 2010.

[41] R. K. Hanusiak. Verificação da autoria de manuscritos com base em atributos genéticos e genéricos da escrita. Dissertação de Mestrado, Pontifícia Universidade Católica do Paraná, 2010.

[42] R. K. Hanusiak, L.S. Oliveira, E. Justino, e R. Sabourin. Writer verification using texture-based features. *IJDAR*, 15(3):213–226, 2012.

[43] R.M. Haralick, K. Shanmugam, e I.H. Dinstein. Textural features for image classification. *IEEE Transactions on systems, man and cybernetics*, 3(6):610–621, 1973.

[44] Z.Y. He e Y.Y. Tang. Chinese handwriting-based writer identification by texture analysis. *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, volume 6, páginas 3488 – 3491 vol.6, Agosto de 2004.

[45] E. C. Herrera-Luna, E. M. Felipe-Riveron, e S. Godoy-Calderon. A supervised algorithm with a new differentiated-weighting scheme for identifying the author of a handwritten text. *Pattern Recognition Letters*, 32(8):1139 – 1144, 2011.

[46] J. H. Holland. *Adaptation in natural and artificial systems*. MIT Press, Cambridge, MA, USA, 1992.

[47] A. Imdad, S. Bres, V. Eglin, C. Rivero-Moreno, e H. Emptoz. Writer identification using steered hermite features and svm. *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, páginas 839–843, Washington, DC, USA, 2007. IEEE Computer Society.

[48] A.K. Jain, R.P.W. Duin, e J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):4–37, 2000.

[49] A.K. Jain, A. Ross, e S. Pankanti. Biometrics: a tool for information security. *IEEE Transactions on Information Forensics and Security*, 1(2):125 – 143, Junho de 2006.

[50] E. Justino. A autenticação de manuscritos aplicada à análise forense de documentos. *TIL - Workshop em Tecnologia da Informação e Linguagem Humana*, páginas 102–106, 2003.

[51] O. Kirli e M. B. Gülmezoqlu. Automatic writer identification from text line images. *International Journal on Document Analysis and Recognition*, páginas 1–15, 2011. 10.1007/s10032-011-0161-9.

[52] J. Kittler, M. Hatef, R. P.W. Duin, e J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[53] B. Ko, S. Kim, e J. Nam. X-ray image classification using random forests with local wavelet-based cs-local binary patterns. *Journal of Digital Imaging*, páginas 1–11. 10.1007/s10278-011-9380-3.

[54] K. M. Koppenhaver. *Forensic document examination: principles and practice*. Humana Press Inc., 2007.

[55] U. Kurban, D. Tursun, A. Hamdulla, e A. Aysa. A feature selection and extraction method for uyghur handwriting-based writer identification. *Proceedings of the 2009 International Conference on Computational Intelligence and Natural Computing - Volume 02*, páginas 345–348, Washington, DC, USA, 2009. IEEE Computer Society.

[56] G. Louloudis, B. Gatos, e N. Stamatopoulos. Icfhr 2012 competition on writer identification challenge 1: Latin/greek documents. *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, páginas 829–834, 2012.

[57] D.G. Lowe. Object recognition from local scale-invariant features. *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, páginas 1150–1157 vol.2, 1999.

[58] S. Al Maadeed, W. Ayouby, A. Hassaine, e J.M Aljaam. Quwi: An arabic and english handwriting dataset for offline writer identification. *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, páginas 746–751, 2012.

[59] U.-V. Marti e H. Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5:39–46, 2002. 10.1007/s100320200071.

[60] U.-V. Marti, R. Messerli, e H. Bunke. Writer identification using text line based features. *Proceedings Sixth International Conference on Document Analysis and Recognition (ICDAR)*, páginas 101 –105, 2001.

[61] J. Martins, Y. M. G. Costa, D. Bertolini, e L. S. Oliveira. Uso de descritores de textura extraídos de glcm para o reconhecimento de padrões em diferentes domínios de aplicação. *XXXVII Conferencia Latinoamericana de Informática*, páginas 637–652, 2011.

[62] J. Martins, L.S. Oliveira, S. Nisgoski, e R. Sabourin. A database for automatic classification of forest species. *Machine Vision and Applications*, 24(3):567–578, 2013.

[63] R. H. C. Melo. Using fractal characteristics such as fractal dimension, lacunarity and succolarity to characterize texture patterns on images. Dissertação de Mestrado, Universidade Federal Fluminense, 2007.

[64] T. Mäenpää. *The Local Binary Pattern Approach to Texture Analysis - Extensions and Applications*. Tese de Doutorado, University of Oulo, 2003.

[65] A. J. Newell e L. D. Griffin. Writer identification using oriented basic image features and the delta encoding. *Pattern Recognition*, 47(6):2255 – 2265, 2014.

[66] A. Nicolaou, M. Liwicki, e R. Ingolf. Oriented local binary patterns for writer identification. Muhammad Imran Malik, Marcus Liwicki, Linda Alewijnse, Michael Blumenstein, Charles Berger, Reinoud Stoel, e Bryan Found, editors, *AFHA*, volume 1022 of *CEUR Workshop Proceedings*, páginas 15–20. CEUR-WS.org, 2013.

[67] R. Niels. *Allograph based writer identification, handwriting analysis and character recognition*. Tese de Doutorado, Donders Centre for Brain, Behaviour and Cognition, Radboud University Nijmegen, Netherlands, 2010.

[68] T. Ojala, M. Pietikäinen, e T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:971–987, 2002.

[69] V. Ojansivu e J. Heikkilä. Blur insensitive texture classification using local phase quantization. Proceedings Image and Signal Processing (ICISP 2008), Cherbourg-Octeville, France, 5099:236-243, 2008.

[70] D. Pavelec, E. Justino, Batista L. V, e L.S. Oliveira. Author identification using writer-dependent and writer-independent strategies. *Proceedings of the 2008 ACM symposium on Applied computing*, SAC '08, páginas 414–418, New York, NY, USA, 2008. ACM.

[71] D. Pavelec, E. Justino, e L. S. Oliveira. Author identification using stylometric features. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 11:59–65, 2007.

[72] E. Pekalska e R. P. W. Duin. Dissimilarity representations allow for building good classifiers. *Pattern Recognition Letters*, 23(8):943–956, 2002.

[73] R. Plamondon e G. Lorette. Automatic signature verification and writer identification – the state of the art. *Pattern Recognition*, 22(2):107–131, 1989.

[74] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. A. Smola et al, editor, *Advances in Large Margin Classifiers*, páginas 61–74. MIT Press, 1999.

[75] D. Puig, M. A. Garcia, e J. Melendez. Application-independent feature selection for texture classification. *Pattern Recognition*, 43:3282–3297, Outubro de 2010.

[76] A. R. Rao e G.L. Lohse. Towards a texture naming system: identifying relevant dimensions of texture. *Proceedings of the 4th conference on Visualization '93*, VIS '93, páginas 220–227, Washington, DC, USA, 1993. IEEE Computer Society.

[77] D. Rivard. Multi-feature approach for writer-independent offline signature verification. Dissertação de Mestrado, École de Technologie Supérieure Univeité Du Québec, 2010.

[78] D. Rivard, E. Granger, e R. Sabourin. Multi-feature extraction and selection in writer-independent off-line signature verification. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(1):83–103, 2013.

[79] H.E.S. Said, K.D. Baker, e T.N. Tan. Personal identification based on handwriting. *Proceedings of Fourteenth International Conference on Pattern Recognition*, volume 2, páginas 1761–1764 vol.2, Agosto de 1998.

[80] S. Santini e R. Jain. Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):871–883, 1999.

[81] K. Saranya e M. S. Vijaya. Text dependent writer identification using support vector machine. *International Journal of Computer Applications*, 65(2):6–11, Março de 2013. Published by Foundation of Computer Science, New York, USA.

[82] K. Saranya e M.S. Vijaya. An interactive tool for writer identification based on offline text dependent approach. *International Journal of Advanced Research in Artificial Intelligence(IJARAI)*, 2(1):33 – 40., 2013.

[83] A. Schlapbach e H. Bunke. Using hmm based recognizers for writer identification and verification. *Ninth International Workshop on Frontiers in Handwriting Recognition, (IWFHR-9 2004)*, páginas 167 – 172, 2004.

[84] A. Schlapbach e H. Bunke. A writer identification and verification system using hmm based recognizers. *Pattern Analysis & Applications*, 10:33–43, Janeiro de 2007.

[85] A. Schlapbach e H. Bunke. Off-line writer identification and verification using gaussian mixture models. Simone Marinai e Hiromichi Fujisawa, editors, *Machine Learning in Document Analysis and Recognition*, volume 90 of *Studies in Computational Intelligence*, páginas 409–428. Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-76280-5-16.

[86] A. Schlapbach, V. Kilchherr, e H. Bunke. Improving writer identification by means of feature selection and extraction. *In Proceedings International Conference on Document Analysis and Recognition (ICDAR)*, páginas 131–135, 2005.

[87] L. Schomaker, K. Franke, e M. Bulacu. Using codebooks of fragmented connected-component contours in forensic and historic writer identification. *Pattern Recognition Letters*, 28(6):719–727, abril de 2007.

[88] L. Schomaker e L. Vuurpijl. Forensic writer identification: A benchmark data set and a comparison of two systems [internal report for the Netherlands Forensic Institute]. Relatório técnico, Nijmegen: NICI, 2000.

[89] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[90] C. Shan, S. Gong, e P.W. McOwan. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[91] I. Siddiqi. *Classification of Handwritten Documents : Writer Recognition.* Tese de Doutorado, Université Paris Descartes, 2009.

[92] I. Siddiqi e N. Vincent. Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features. *Pattern Recognition*, 43:3853–3865, Novembro de 2010.

[93] M. Sreeraj e S. M. Idicula. A Survey on Writer Identification Schemes. *International Journal of Computer Applications*, 26(2):23–33, julho de 2011.

[94] S. N. Srihari, S. Cha, H. Arora, e S. Lee. Individuality of handwriting. *Journal of Forensic Sciences*, 4:47, 2002.

[95] M. Unser. Sum and difference histograms for texture classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(1):118–125, Jan de 1986.

[96] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[97] J. F. Vargas, C. M. Travieso, J. B. Alonso, e M. A. Ferrer. Off-line signature verification based on gray level information using wavelet transform and texture features. *International Conference on Frontiers in Handwriting Recognition*, 0:587–592, 2010.

[98] J. Yang e V. G. Honavar. Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems*, 13(2):44–49, março de 1998.

[99] E. N. Zois e V. Anastassopoulos. Morphological waveform coding for writer identification. *Pattern Recognition*, 33(3):385 – 398, 2000.